URN: urn:nbn:de:kobv:b4-opus-24346

SERGE ROSMORDUC,
The Ramses project in perspective. Managing evolving linguistic data,

in:
Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011, herausgegeben von Ingelore Hafemann, Berlin 2013, S. 109-120.

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Thesaurus Linguae Aegyptiae 4

Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011

herausgegeben von Ingelore Hafemann

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Thesaurus Linguae Aegyptiae

4

BERLIN 2013

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Perspektiven einer corpusbasierten historischen Linguistik und Philologie

Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011

herausgegeben von Ingelore Hafemann

BERLIN

2013

Dieser Band wurde durch die gemeinsame Wissenschaftskonferenz im Akademienprogramm mit Mitteln des Bundes (Bundesministerium für Bildung und Forschung) und des Landes Berlin (Senatsverwaltung für Wirtschaft, Technologie und Forschung) gefördert

URN: urn:nbn:de:kobv:b4-opus-24310

INHALTSVERZEICHNIS

VORWORT 7

GREGORY CRANE & ALISON BABEU
Global Editions and the Dialogue among Civilizations 11

HISTORISCHE CORPUS-PROJEKTE – SYNCHRON UND DIACHRON

STÉPHANE POLIS & JEAN WINAND
The Ramses project. Methodology and practices in the annotation of Late Egyptian Texts 81

SERGE ROSMORDUC
The Ramses project in perspective. Managing evolving linguistic data 109

DIETER KURTH
Das Edfu-Projekt. Ziel, Methode und Verarbeitung der lexikographischen Ergebnisse 121

INGELORE HAFEMANN & PETER DILS
Der Thesaurus Linguae Aegyptiae – Konzepte und Perspektiven 127

GÜNTER VITTMANN
Zur Arbeit an der Demotischen Textdatenbank: Textauswahl 145

GERNOT WILHELM
Das Hethitologie Portal Mainz 155

JOST GIPPERT
The TITUS Project. 25 years of corpus building in ancient languages 169

KURT GÄRTNER & RALF PLATE
Die Doppelfunktion des digitalen Textarchivs als Wörterbuchbasis und als Komponente der Online-Publikation. Am Beispiel des Mittelhochdeutschen Wörterbuchs 193

HANS-CHRISTIAN SCHMITZ, BERNHARD SCHRÖDER & KLAUS-PETER WEGERA
Das Bonner Frühneuhochdeutsch-Korpus und das Referenzkorpus ‚Frühneuhochdeutsch' 205

ALEXANDER GEYKEN
Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv 221

BRYAN JURISH
Canonicalizing the Deutsches Textarchiv 235

WORTGESCHICHTE - TEXTGESCHICHTE - SPRACHGESCHICHTE: TRADITION UND INNOVATION BEI DER TEXTPRODUKTION

FRANK FEDER & SIMON D. SCHWEITZER
Auf dem Weg zu einem integrierten Lexikon des Ägyptisch-Koptischen 245

FRIEDHELM HOFFMANN
Die Demotische Wortliste – virtuell erweitert 263

GÜNTER VITTMANN
Kursivhieratische Texte aus sprachlicher und onomastischer Sicht 269

MATHEW ALMOND, JOOST HAGEN, KATRIN JOHN, TONIO SEBASTIAN RICHTER & VINCENT WALTER
Kontaktinduzierter Sprachwandel des Ägyptisch-Koptischen: Lehnwort-Lexikographie im Projekt Database and Dictionary of Greek Loanwords in Coptic (DDGLC) 283

THOMAS GLONING
Historischer Wortgebrauch und Themengeschichte. Grundfragen, Corpora, Dokumentationsformen 317

LOUISE GESTERMANN
Die altägyptischen Sargtexte in diachroner Überlieferung 371

THOMAS STÄDTLER
Überlegungen zu Textsorte und Diskurstradition bei der Beschreibung von Textcorpora und ihr Bezug zur lexikographischen Forschung 385

VORWORT

Die internationale Tagung „Perspektiven einer corpusbasierten historischen Linguistik und Philologie“ vom 12. – 13. Dezember 2011 am Akademienvorhaben „Altägyptisches Wörterbuch“ der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) war dem Thema des Aufbaus und der Nutzungsperspektiven elektronischer Textcorpora und Wörterbücher in den historischen Sprachen gewidmet. Die Teilnehmer, Vertreter der Ägyptologie, der Hethitologie, Indogermanistik sowie Referenten aus der historischen Lexikographie des Mittel- und Frühneuhochdeutschen und des Altfranzösischen diskutierten vor allem über die Veränderungen, die mit dem Einsatz elektronischer Erfassungs- und Verarbeitungsprozeduren einhergehen. Vertreter der Computerlinguistik vom „Zentrum Sprache“ der BBAW wurden in die Diskussionen einbezogen. Dort beschäftigt man sich seit Jahren mit dem Aufbau großer elektronischer Textcorpora (DWDS), darunter auch solcher, die historische Texte (DTA) für die elektronische Nutzung ermöglichen.

Die größte Herausforderung dieser neuen elektronischen Corpora und Wörterbücher ist es, sowohl den Methoden und damit den wissenschaftlichen Ansprüchen der traditionellen Philologie und Lexikographie unbedingt verpflichtet zu bleiben als auch neue Gebiete wie die Corpus- und Computerlinguistik für die historischen Sprachen zu öffnen. Die Teilnehmer haben gemeinsam und disziplinenübergreifend die Möglichkeiten und Grenzen der Datenerfassung, ihrer Präsentation und den Nutzen neuer Auswertungsprozeduren diskutiert.

Unter dem ersten Thema „Historische Corpusprojekte – synchron und diachron“ wurden elektronische Corpora vorgestellt und ein intensiver Austausch darüber geführt, welche Datenstrukturen die linguistischen Inhalte in adäquater Weise abbilden. Wichtig war die Frage, auf welche Resonanz diese elektronischen Corpora bei den Nutzern gestoßen sind und welche Erwartungen und Anforderungen aus den verschiedenen Fachdisziplinen an die Projekte herangetragen werden. Der Austausch über Nutzungsperspektiven elektronischer Corpora schloss auch die Diskussion über die Erarbeitung projektübergreifend einsetzbarer Standards der Codierung und Strukturierung historischer Textdaten mit ein. Hinsichtlich einer mittel- und langfristigen Nutzbarkeit sowie einer langfristigen Datensicherheit stehen solche Fragen zunehmend im Focus und einige aktuelle Initiativen dazu wurden vorgestellt. Spezielle technische Aspekte

elektronischer Datenerfassung und automatischer Analyse- und Speicherungsverfahren elektronischer Textdaten konnten am letzten Tag als ein Themenschwerpunkt mit den Programmierern diskutiert werden.

Ein zweiter Schwerpunkt waren konkrete Fragstellungen aus der historischen Lexikographie und diachronen Textanalyse. Für das Ägyptische ist der diachrone Ansatz auf Grund der über viertausendjährigen Textüberlieferung von großer Relevanz. Themen wie historischer und/oder textgattungsspezifischer Wortgebrauch, die Erarbeitung diachroner Wortlisten und Aspekte des kontaktindizierten Sprachwandels konnten disziplinübergreifend zwischen den Ägyptologen und den Kollegen der historischen Lexikographie des Mittel- und Frühneuhochdeutschen und des Altfranzösischen behandelt werden.

Mit dem Abendreferenten Gregory Crane, dem Begründer der „Perseus Digital Library", wurde ein breites Publikum angesprochen. In seinem Vortrag hat er noch einmal die hohe Relevanz und die neuen Möglichkeiten der Einbeziehung zahlreicher Wissenschaftler und einer interessierten Öffentlichkeit in die Projektarbeit demonstriert, die das Internet auf völlig neue Weise eröffnet hat. Die Herausgeberin ist sehr froh, seinen programmatischen Beitrag zu diesem Thema, dessen schriftliche Form er gemeinsam mit Alison Babeu erarbeitet hat, ebenfalls in diesem Band präsentieren zu können.

Wir danken der Berlin-Brandenburgischen Akademie der Wissenschaften für die umfassende Unterstützung unserer Projektarbeit und ganz speziell der Vorbereitung dieser Konferenz sowie der Möglichkeit, die Akten auf dem E-Doc-Server der Akademie veröffentlichen zu können.

Der Hermann und Elise geborene Heckmann Wentzel-Stiftung sei hiermit ausdrücklich für die unbürokratische und großzügige finanzielle Unterstützung dieser erfolgreichen Tagung gedankt.

Das Akademienvorhaben „Altägyptisches Wörterbuch" konnte sich als aktives Mitglied des Weiteren auf das „Zentrum Grundlagenforschung Alte Welt" stützen, dem alle altertumswissenschaftlichen Vorhaben der BBAW angehören. Dem Zentrum ist es zu danken, dass der Abendvortrag von Gregory Crane einem breiteren Publikum dargeboten werden konnte.

Allen Autoren dankt die Herausgeberin für ihre anregenden Diskussionen und die qualitätvollen Beiträge in diesem Band.

Auf eine Gesamtbibliographie wurde verzichtet und die Abkürzungen der in den ägyptologischen Beiträgen erwähnten Zeitschriften und Reihen folgen dem Lexikon der Ägyptologie, herausgegeben von Wolfgang Helck und Wolfhart Westendorf, Band VII: Nachträge, Korrekturen, Indices, Wiesbaden 1992, XIV-XIX.

Ganz besonders sei schließlich Frau Angela Böhme für die gewissenhafte redaktionelle Bearbeitung der Manuskripte gedankt sowie Dr. Simon Schweitzer für seine Hilfe beim Erstellen des Layouts.

Berlin, Mai 2013 Ingelore Hafemann

THE RAMSES PROJECT IN PERSPECTIVE
MANAGING EVOLVING LINGUISTIC DATA

SERGE ROSMORDUC

1. Introduction

Back in 2006, when we started the Ramsès Project [WINAND *et al.* (2008), ROSMORDUC *et al.* (2008)], our primary concern was to develop an efficient tool for data input, and to deliver it to the encoding team as soon as possible. We had to work on a tight schedule, and, keeping in mind ease of use, and linguistic coherence as prime requisites, we did not have time to handle fully the problem of the evolution of our data and of our own view of the data.

Needless to say, we have paid the price for this a number of times, and, in more than one case, we had to perform large-scale modifycations of the data, which usually required both automated and manual processing. Most of the time, those modifications involved lemma or parts of speech modifications, for which our system is rather versatile. In a few instances, we did modify the very structure of our lexicon (that is, the way the database tables are organized). Even if, in retrospect, no major problem occurred, the process was quite taxing.

As the initial phase of development of Ramsès is almost done, with a working prototype of a syntactic editor, we have started to think about ways of improving the encoding process, and securing our data consistency. This paper explains the current state of our ideas on the subject.

2. A short typology of changes in Ramsès

We will start by looking how data evolves in the "Ramsès" ecosystem, and try to roughly classify those changes. Doing so will probably amount to beating a dead horse, as most similar databases will have met the same problems.

Anyway, a short description of the organisation of the Ramsès database is necessary to explain the extent of our problems. In Ramsès, the structure for representing the texts is quite complex. First, each entry is documented both in terms of content (a text, which might be known from multiple copies, and which has a number of characteristics, such as a particular blend of language

(from Middle Egyptian to full Late Egyptian) or genre of text. The entries are also documented in terms of original documents, which have they own characteristics (date, writing support or script for instance).

The content of the text is a sequence of lemma analysis, which record word spelling, lemma and inflexions. Actually, the spellings, lemma and inflexions are recorded in a lexicon, and a text content is a list of references to the said lexicon.

In this respect, Ramsès is quite different from non-lemmatised text databases. In a non lemmatised database, the information in a particular entry is relatively stand-alone, and quite impervious to changes in the rest of the base. In Ramsès, the data in a given entry depends heavily on the lexical database.

This being said, we can now categorise changes in Ramsès in two families. The first kind of change is due to modifications of data, which can occur at multiple points. This is the most mundane type of changes, although, in the current state of the database, it can have huge consequences in some cases.

The second and most drastic change is structural. In some cases, one can consider modifying the structure of the database itself.

2.1 Data Modification

Data modification can alter the texts, the lexicon, and, in Ramsès, even the lexicon structure. In normal edition mode, the Ramsès encoders will create and modify text entries routinely. They will also need to enter new words, new spellings and new inflexions, thus modifying the lexicon. Let's consider a few examples, and the problems they incur.

The first example is *simple text modification*. An encoder can modify an existing text, adding, deleting or replacing part of its content. Our current system does not remember the modifications made to a text and, thus, present only the latest stage of editing. However, it is quite desirable to keep a history, for a variety of reasons, especially if text edition is in part collaborative. It can allow to reverse bad manipulations, like accidental erasures of parts of the text.

A second type of modifications is *lexical modification*. In Ramsès, all encoders can add, delete, or modify lemmas, inflexions and spellings in the lexicon. Obviously, a change in a given lemma (for instance, a change in its transliteration) will have large consequences on *all* the texts which use it. It is quite important to be able to review

those changes and to be able to cancel them if needed. Even the creation of a lemma can be a problem, for the said lemma can be already encoded (especially if we consider the variability of Late Egyptian orthography).

Finally, a large number of features of the lexicon are encoded as data, and can be modified by the administrators of the system. For instance, parts of speech, the inflexions associated with them, and their attributes can be modified (which would allow, for instance, to use the database to create a purely Middle Egyptian database). Those modifications are quite critical, for they potentially affect, directly or indirectly, large segments of the database. Let's take a few examples.

Creating a new part of speech, adding information to existing parts of speeches or inflexions is not a simple addition. It can affect encoded texts, in that their encoding might become partly obsolete.

For instance, if the database contains only one form for the infinitive, and if we decide to distinguish *status absolutus*, *status constructus* and *status pronominalis*, and introduce this new distinction in the lexicon, all texts encoded prior to the addition will become obsolete and need reviewing (or, at the very least, will need to be specifically marked).

Other modifications can call for more drastic text revisions. For instance, when we added the notion of "phrase determiners", to indicate the determinative which can occasionally appear after a relative clause or a complex noun phrase, a number of texts had already been encoded, in which those determinatives had been somehow artificially related to the last word of the noun phrase. We had to change the segmentation of the texts in a few other cases, in particular when, after many discussions, we decided that most compound words (and in particular titles) were to be analysed when it made sense and not encoded as a unit (*ḥry.w-š* is one word, but *ḥry pḏ.t* is encoded as two units).

2.2 Structural Modifications

In extreme cases, the core structure of the database itself can change. For instance, the spellings were initially dependent of lemma and inflexions. We had an entry "𓅱𓄿" as spelling for the circumstantial *jw*, and another entry with the same glyphs for the unusual spelling of the "r" preposition. At some point, it appeared that we really needed to consider spellings as independent entities which could be shared between lemma and inflexions: sometimes there is no doubt upon a spelling, but the corresponding lemma is unclear, for instance. In this

case, we had to change the structure of the database tables, and to correct all the texts. It was however possible to do it through an automated process.

In the near future, a number of important structural changes will probably take place. In our current system, we have used a number of hierarchical thesauri for texts descriptions: dating, writing system, geographical information use a thesaurus. However, the lexeme descriptions do not. In a number of cases, it might be interesting to allow some kind of inheritance between lexical attributes. For instance, substantives have a number of attributes like "animate", "proper name", "god"... and, obviously, it would be better to consider "proper name" as a sub-type of "animate". Inflexions could benefit from the same system: a meta-category "*sḏm=f*", could be considered as a super-category for perfective and subjunctive Late Egyptian *sḏm=f*, avoiding the current problem in the prologue of many royal texts, where it is often difficult to choose between the two forms.

A last kind of structural change which will take place at some point is that the lemma themselves need to be hierarchically organised, with links to their roots, meaning classification, etc... as it is now the case in the *thesaurus linguae aegyptiae*. When this is done, it is quite likely that parts of the texts will require a revision.

Fortunately, most of those large, structural changes, can be *partially* automated. Actually, we would probably not think of doing them if it wasn't the case. However, changes made to such an extent mean drastic modifications, and, hence, we have sometimes considered necessary changes with reluctance.

2.3 Temporary Conclusions

To summarise our current discussion, the core of the problem is that our data is precious. It amounts to thousands of hours of work. We don't want to loose part of it due to an incorrect manipulation. Another point is that, as a collaborative work, we wish it to be coherent. As the world is not perfect, the system must be able to deal with semi obsolete data (if we delete a lemma, we probably don't want to delete the texts which use it altogether).

We want to know why the database changes and how it changes. This will allow us to cancel erroneous modifications, and to understand systematic errors too.

3. Managing data history

A text database like Ramsès is a collaborative project. It is rather structured, with a validation process for texts, but still, even if it is not a Web 2.0 application, we do have a significant team of encoders, and the decisions of each one is important and impact the others (especially as far as the lexicon is concerned). In this section, we are going to consider the options for monitoring the data changes.

3.1 State of the Art

The first possible solution would be to centralise most decisions. One or two experts would need to approve some operations (like lemma creation) before they are done. However, this would not solve all the problems, as even experts can change their minds, or mis-use the user interface.

Still on an organisational level, some projects have introduce interesting practices: in the *Syntactic Reference Corpus of Medieval French* [PRÉVOST, S. & A. STEIN (2012:6)], a text is encoded twice independently. For each entry where there is a disagreement, a discussion takes place between the two encoders, and, if needed, a senior encoder takes the final decision choice. However, this improve the initial content of the database, but does not constitute a management tool for existing entries.

The Open Richly Annotated cuneiform corpus (http://oracc.museum.upenn.edu), uses the notion of versions of the database. This is quite interesting, because if someone quotes the database for a given result, the quotation can be made obsolete by modifications of the texts. Versioning allows consistent quotation, and might even allow to reproduce searches at a given point in time.

The IDP project (http://papyri.info), which regroups the Duke Databank, the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV), and the Advanced Papyrological Information System (APIS), has used systems like *subversion* and *git* [BODARD *et al.* (2011)] for managing their text database history. *Subversion* and *git* have been developed to manage software projects. In large development teams, it is rather usual that a number of programmers modify the same files, and version control systems like these ones save a precise history of modifications and help to keep a consistent view of the code, reverting changes if needed.

A number of other systems can be considered, such as wiki oriented ones, which stores a precise history of each page.

A last possibility, which is not currently used by textual databases (as far as we know), is to keep a precise history, not of the *text content*, as is done explicitly or implicitly by the previous systems, but of *operations made on the database*. Actually, this is what most softwares use in-memory to implement "undo/redo" facilities.

3.2 History Management and the needs of Ramsès

One peculiarity of Ramsès, as compared with the databases listed above, is that it has a very structured lexicon. A text in Ramsès is basically a list of references to its lexicon. The lexicon itself cannot be considered as a text, and the granularity of actions taken on it needs to be very fine. In other words, it will probably be enough to store a number of versions of each text in the base (we might even consider a policy where only a few versions would be kept indefinitely). But, on the other hand, when we come to the lexicon, we must be much more precise. For instance, some operations on the lexicon are much more sensitive than others. Adding a new spelling is something relatively innocuous. On the other hand, removing a lemma from the lexicon has potentially huge consequences on existing texts. Indeed, the importance of a change in the database could be measured by considering how many modifications it would entail in the rest of the data. In a similar way, adding a new attribute to a lemma entry is not as serious as changing its part of speech.

The solution to this problem is to record the *operations* on the database: creation of an entry, modifications of an attribute, deletion or an entry. In theory, starting with an empty set of data, and having the whole history, we would be able to re-create the data as seen at any date. In practice, the database must contain two different parts: on one hand, a large history, recording for each modification its date, its author, and the data needed to perform (and maybe to undo) the modification. On the other hand, a "view" of the current state of the database content is needed for efficient access.

An interesting property of this model is that it allows to examine and eventually to cancel some operations selectively. For instance, attribute changes are not all equal. With a precise list of operations, it is possible to review changes in lemma parts of speech (and only them), and to cancel them *without cancelling the rest of the work*. That is, if a particular change is thought to be erroneous, we don't need to revert the whole database to a state anterior to the change. We can compute precisely what modifications are dependant on the change, and revert them. It can even be done on a rather fine level. For in-

stance, we could allow links between a lemma and a spelling to be done, while cancelling a part of speech change on the lemma.

In the future, we want a better control on user modifications. Most of this control will be done up-front (that is, some modifycations won't be available to all encoders in the first place). However, the history management will allow us to catch even the problems we haven't foreseen, and provide a safer environment even for expert encoders.

Last, but not least, the reversibility of all operations will be a welcome addition when performing large (and semi-automated) changes, alleviating the burden on the software engineer performing the modifications (who has usually been the present author until now).

3.3 The status of Deletion

Among all modifications, deletion has a special status. If we delete a reference to a lemma, spelling or document, what should happen to the texts which use it? Even when the deletion is "correct", if we perform it blindly, we will lose information. Suppose for instance that we had created a "ghost" lemma, which should be merged with others (consider for instance *snj* vs. **sš* in Middle Egyptian dictionaries). If the "ghost" lemma is actually deleted from the database, all references to it in the database texts will be point nowhere, and be difficult to amend. Thus plain deletion is not a good idea.

The solution is simple: we should not actually delete data, but mark it as "obsolete". Obsolete entries will be kept, but it will be impossible to select an obsolete lemma when typing a new text, for instance. Creating a list of obsolete entries and their uses is quite easy, and the database administrators will then be able to rectify all occurrences of the virtually deleted data.

3.4 Final Words on History Management

The only real problem with this approach is that the size of the database might grow a bit too much. It is not sure that such fears are founded, as they depend on the actual behaviour of encoders. Actual data is needed there. In any case, it is possible to imagine some clean-up phases, where some part of the database history will be forgotten.

Finally, a last interesting feature of the process we want to use in the future is that keeping an exact history of what happens is also a way to improve our understanding of the encoding process itself.

Patterns of operations made by encoders can be found, and used to improve the interface. Common errors can be detected more easily and prevented, etc.

4. Managing Structural Change

Keeping a log of changes works well for data modifications. However, at least parts of our database are liable to structural modifycations. How can we deal with them? We can use an age-old technique of computer scientists for this: transform structure into data. This is more or less the approach used by the Notabene system [MAZZIOTTA (2010)].

4.1 Graph databases

In recent years, for reasons quite alien to ancient languages databases, a family of software globally called "NoSQL" databases has become fashionable. SQL relational databases are very efficient and very secure for well structured, well understood data. But they are not very well suited for changing, distributed and semantic data. Typically, keeping information on web pages, managing the data for very large but loosely structured databases (like those used by *tweeter* or *facebook*), is not a task where classical relational databases perform well.

Huge distributed databases are not really an issue for ancient languages projects, as even an exhaustive database of Egyptian would be very small if we compare it with, for instance, a newspaper archive. But the relative freedom they offer in terms of modelling is interesting.

Databases like Neo4J or OrientDB are built on graph theory. The same theory stands behind the semantic web and the *Resource Description Framework* [MANOLA & MILLER (2004)]

In these systems, there are basically two kinds of elements: "*nodes*", and "*links*" between nodes.

If we consider, form instance, inflexions and lemma in our current (relational) system, we have a "lemma" table and an "inflexion" table. Saying that entry 123332 is a lemma, with an inflexion with id 56562 means in our current system that there is an entry with id 123332 in the lemma table, and that the entry with id 56562 in the inflexions table has a foreign key attribute pointing to 123332.

If we ever decide to change the architecture, we need to modify the database table structure, which is a heavy operation. Keeping a history of the operations done would not be simple, either.

Now, with a graph database, all of 123332, 56562, "lemma" and "inflexion" would be nodes. A "is-a" link would be used to specify the class of each node (Figure 1). That is, both data and structure would use the same formalism.

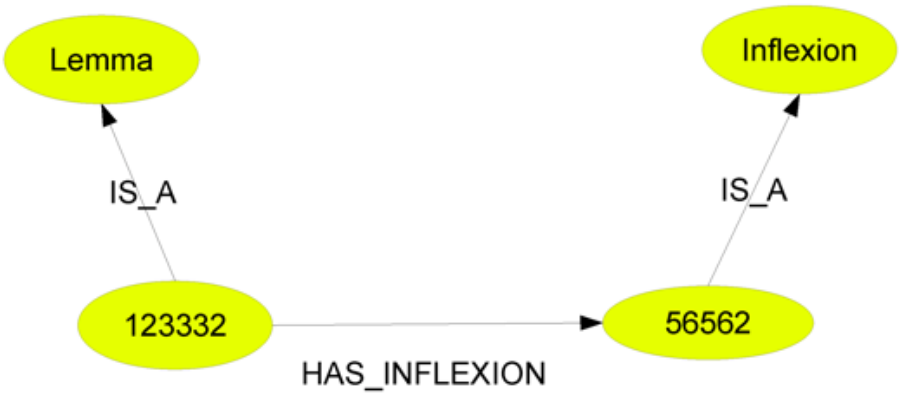

Figure 1: graph database example

The advantage of this representation is that it is very uniform. It is possible to accommodate changing structures with it, and to keep for some time parts of an old model alongside a new one.

Another interesting element is that it can be easily coupled with a history system. There is a small set of possible modifications: node, link, and attribute creation, deletion or modification. By recording them, one gets a history which covers all possibilities on a database.

The downside of such a general system is that it's more complex to manage and to interrogate than an SQL database. However the current version of Ramsès works with the whole data in-memory, and the database itself is only a convenience to store the information on a permanent basis.

4.2 The Case of Syntax

The Ramsès database is planned to contain ultimately a fully parsed syntactic annotation of all the texts. The problem is that, whereas lexical annotation is relatively stable, our syntactic formalism will certainly evolve a lot. The way we deal with it currently might be seen as a forerunner of the general system to come. We have decided to represent the tagging system for syntax as a "loose grammar", which states what kinds of phrases are available, and how they can

relate. The grammars are described as simple text, in an *ad-hoc* formalism, which is described in POLIS *et al.* (2013); the idea of keeping an external, user-created and explicit definition of the grammar is taken from N. MOZZIOTA (2010).

The encoder analysing a text needs to choose which grammar he will use, and the system ensures the analysis respects the chosen system.

The system is relatively rich: it's possible to create syntactic categories, to assign them attributes, to state which kinds of groups are fit to fulfil a particular function in a parent group (for instance that the subject is supposed to be a noun phrase). Provision is also made for non hierarchical inter-group links, which can occur for instance between a pronoun and its antecedent, etc.

As an example, the following fragment of grammar describes a simple annotation scheme, which defines four kinds of syntactic groups: *groups*, *noun phrases*, *adverbial phrases*, and *propositions*. A simple inheritance mechanism allows us to factor common attributes. The "group" can bear common attributes, in our present case, comments (which are free text).

The noun phrase has a specific attribute, *defined*, whose possible values are declared in the "TYPE" declaration at the beginning of the file. "Definiteness" can be either "unset", "defined", "undefined", or "doubtful".

The last interesting feature in this file is the definition of "proposition", in which we define two possible children: subject, which must be a noun phrase, and adjunct, which must be an adverbial phrase.

```
ANNOTATION SCHEME "st_1"
TYPE definiteness ENUM unset defined undefined
doubtful ENDTYPE
GROUP group
  ATTR comment TEXT * ENDATTR
ENDGROUP
GROUP nounPhrase EXTENDS group
  ATTR defined definiteness ONE unset ENDATTR
ENDGROUP
GROUP adverbialPhrase EXTENDS group
ENDGROUP
// proposition
GROUP proposition EXTENDS group
```

```
  CHILD subject CHILDTYPE nounPhrase ENDCHILD
  CHILD adjunct CHILDTYPE adverbialPhrase ENDCHILD
ENDGROUP
```

5. Conclusion

The next step for us is to create an experimental implementation of the graph database. It will be a good stepping stone for programming an historical log. Meanwhile, we will try to generalize the database. The software behind Ramsès (excluding libraries such as Jsesh) is already 80000 lines long, and it would be nice if it could benefit other projects in the near future.

6. BIBLIOGRAPHY

BODARD, G. *et al.*, 2011: Lessons from the conversion of the Duke Databank of Documentary Papyri from legacy formats into EpiDoc TEI XML, abstracts of the Digital Humanities 2011 conference, Stanford, in: https://dh2011.stanford.edu/wp-content/uploads/2011/05/DH2011_BookOfAbs.pdf, p. 31.

MANOLA, F. & E. MILLER, 2004: *RDF Primer*, http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

MAZZIOTTA, N., 2010: Logiciel NotaBene pour l'annotation linguistique. Annotations et conceptualisations multiples, in: *Recherches qualitatives. Hors-série "Les actes"*, 83-94.

POLIS, S. & S. ROSMORDUC, 2013: Building a construction-based treebank of Late Egyptian. The syntactic layer in Ramsès, in: POLIS, S. & J. WINAND (eds.), *Texts, Languages & Information Technology in Egyptology. Selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptologie), Liège, 6-8 July 2010*, Ægyptiaca Leodiensia 9, Liège, 45-59.

PRÉVOST, S. & A. STEIN, 2012: *Syntactic Reference Corpus of Medieval French et l'ordre des compléments du verbe en ancien français*, Séminaire "Lectures en linguistique experimentale", Université Paris 7, http://www.uni-stuttgart.de/lingrom/stein/downloads/prevost-stein-objets-afr-handout.pdf.

ROSMORDUC, S. *et al.*, 2008: Ramsès, a new Research Tool in Philology and Linguistics, in: STRUDWICK, N. (ed.), *Information Technology and Egyptology in 2008. Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique et Egyptologie), Vienna, 8-11 July, 2008*, Piscataway, NJ, 155-166.

WINAND, J. *et al.*, 2008: Ramses. An Annotated Corpus of Late Egyptian, in: KOUSOULIS, P. (ed.), *Proceedings of the Xth IAE Congress, Rhodos, 2008*, (in press).

OrientDB: http://www.orientdb.org/.

Neo4J: http://neo4j.org/.