Title
Computational Linguistics in Egyptology

Permalink
https://escholarship.org/uc/item/0fk4n4gv

Journal
UCLA Encyclopedia of Egyptology, 1(1)

Author
Rosmorduc, Serge

Publication Date
2015-04-02

Copyright Information

Peer reviewed

COMPUTATIONAL LINGUISTICS IN EGYPTOLOGY

علم اللغة الحاسوبي (الآلي) وعلم المصريات

Serge Rosmorduc

EDITORS

WILLEKE WENDRICH
Editor-in-Chief
University of California, Los Angeles

JACCO DIELEMAN
Editor
University of California, Los Angeles

ELIZABETH FROOD
Editor
University of Oxford

JOHN BAINES
Senior Editorial Consultant
University of Oxford

JULIE STAUDER-PORCHET, ANDRÉAS STAUDER
Area Editors Language, Text and Writing
Swiss National Science Foundation & University of Basel

COMPUTATIONAL LINGUISTICS IN EGYPTOLOGY

علم اللغة الحاسوبي (الآلي) وعلم المصريات

Serge Rosmorduc

Verarbeitung natürlicher Sprache für Ägyptologie
Informatique linguistique en égyptologie

Computer-assisted approaches to text and language, referred to as computational linguistics, represent a developing field in Egyptology. One of the main concerns has been and continues to be the encoding of hieroglyphic signs for computers. The historical standard in this respect is the Manuel de Codage; a Unicode encoding has also been recently developed. Computer-assisted approaches also provide helpful tools notably for creating, annotating, and exploiting text databases. After pioneering work in the 1960s, a number of large text databases have been developed since the 1990s, for example, the Thesaurus Linguae Aegyptiae *or the "projet Ramsès." Ongoing projects involve automated text processing and analysis for Egyptian, especially automated transliteration, part-of-speech tagging, and optical character recognition.*

يشار إلى استخدام منهجية المساعدة – الحاسوبية فيما يختص بالنص واللغة علي انه علم اللغة الحاسوبي (الآلي) ، وهذا يمثل حقل متطور في علم المصريات. كانت واستمرت عملية ترميز (تكويد) العلامات الهيروغليفية لأجهزة الكمبيوتر احد الاهتمامات الرئيسية، المعيار التاريخي في هذا الصدد هو "Manuel de Codage" وهو عبارة عن نظام ترميز يونيكود تم تطويره مؤخرا. كما توفر منهجية المساعدة – الحاسوبية أدوات مفيدة خصوصا لإنشاء، وتذييل، واستغلال قواعد البيانات النصيه. بعد العمل الرائد في 1960 تم تطوير عدد من كبير قواعد البيانات النصية منذ 1990، على سبيل المثال، Thesaurus Linguae Aegyptiae أو "مشروع رمسيس". المشاريع الجارية تنطوي علي التجهيز الآلي للنصوص وتحليلها من أجل اللغه المصريه، بالاخص الترجمة الصوتية الآلية، وتصنيف اقسام الكلام.

Computational linguistics, also referred to as Natural Language Processing (probably with a more restrictive meaning), is a large and heterogeneous field covering the processing and representation of linguistic data with the help of computers. The processing can be based on linguistic or statistical grounds. Born in the guise of "automated translation" in the early 1960s, the field has undergone dramatic changes, with a revolution in the late 1990s when the advent of the WWW lead to an emphasis on huge text databases. At the same time, the need to "enrich" the texts with annotations provided the impetus for numerous works on text representation. These resulted in a number of standards proposed by the Text Encoding Initiative (TEI) and embodied, for philology, in the epidoc suite of software (EPIDOC).

The relevance of computer linguistics for Egyptology is twofold. First, the work on document representation and structure, and the advent of collaborative tools to create such documents, can be seen as a rationalization of

traditional practices. Annotating texts, quoting colleagues' works, comparing variants, etc. are old-time scholarly practices for which the non-sequential navigation, natural in a hypertext system, is particularly suitable. Second, creating, annotating, and using text databases to extract meaningful information, be it linguistic or cultural, is not an easy task. In this case, natural language processing can provide helpful systems. Since the early 1990s, a large part of the work in these fields has been devoted to automated tools. These are not supposed to mimic the human processing of texts, but should try to be statistically accurate enough to remove most of the tedious work from the human operator.

The use of computers to store and process ancient texts can be traced back to 1948, when Father Roberto Busa, with the help of IBM, started an index of the works of St. Thomas Aquinas. In Egyptology, the first attempts can be dated to the end of the 1960s, with the works of W. Schenkel, on the one hand, and the early versions of Glyph by J. Buurman, on the other. Natural language processing applied to ancient languages has been the subject of a number of recent publications and conferences (Denooz and Rosmorduc 2009; Piotrowski 2013).

Text Representation

Currently, there are well established standards only for encoding hieroglyphic texts (and thus hieratic texts by transcribing them in hieroglyphs). Regarding Demotic, existing databases use transliteration as encoding, following the main trend of Demotic studies. Coptic texts now have a proper Unicode encoding and can thereby be processed as all alphabetic writing systems. In encoding hieroglyphic texts, two different problems must be solved. The first is to list the individual signs; the second is to lay out the signs in such a way as to reproduce the original texts. While most existing systems do both, the problems are discussed separately here.

Sign Encoding

While the problem of sign encoding is as old as Egyptological typography (Janssen 1972), it became acute with the advent of the first computer systems. Hieroglyphic transcription entails choices in a continuum between facsimile and normalization. As is evidenced, for instance, in Sethe's publication of the Altaegyptische Pyramidentexte, these choices are often made at the level of individual signs, not of whole texts. This can make full text searches difficult (Rosmorduc 2002).

The Manuel de Codage and the Extended Library

The 1980s saw much work on the subject (Baines and Griffin 1988; Hallof 1988; Stief 1988; Tiradritti 1988; Meeks 1995), and a standard called the Manuel de Codage (abbr. MdC; Buurman 1988) eventually emerged from the works of the "informatique et égyptologie" working group. The proposed encoding was based on the Gardiner sign lists, with extensions to cover parts of the IFAO fonts. In this original list, "there is a clear distinction made between graphemes and graphics variant" (Buurman 1988: 51), at least in the intent. The list was later extended to cover specific publication needs and became the extended library of Winglyph and Mac Scribe. While full text databases were explicitly considered in the first standards, this aspect remained underused for a long time, a notable exception being the works of J. Hallof and H. van den Berg (Hallof and van den Berg 1992), with, for instance, the creation of the index of the Osirian chapels in Dendera (Cauville 1997). Due to user demands, most extensions were driven by practical printing purposes.

Currently, the main problem of the list is the lack of published documentation for signs. For text database purposes, the proposed list sometimes lacks homogeneity. For instance:

R49 , as a variant of R10 , should probably be encoded as R10H (as is R10A). Furthermore, some signs have a huge number of graphical variants (thus, 27 variants

in the case of R3), which probably do not even cover the whole spectrum of possibilities, and thus give a false impression of palaeographic fidelity.

Another urgent problem to solve is the absence of a centralized portal to code "new" signs, which could result in incompatible sign codes being given by font creators. Yet, in spite of these growth problems and of its various dialects, the Manuel de Codage has nonetheless provided the community with a standard, which has been used for more than 25 years (Polis and Rosmorduc 2013).

Unicode

Unicode is a standard encoding for characters, i.e., individual meaningful graphemes. In the Unicode terminology, the character is seen as an abstraction, which gives linguistic information regardless of its actual graphical realization, which is called a glyph. Thus, the character "a" is the character "a" regardless of the actual shape used to display it. In this respect, both Y1 ⎯ and Y2 ⎯ should be regarded as two glyphs for the same character. Encoding a character means giving it a numeric code, which is then used to represent this character in memory; for instance, "a" has the code "97" in Unicode. In theory, Unicode's main concern is how to represent the texts' information, and not how to display it. In practice, this is a little blurred by the presence of so-called "legacy characters," for compatibility with existing fonts, as well as by the limitations of current software and the fact that the character/glyph distinction is not always as clear as one would like it to be. For Egyptology, Unicode applies to two domains: encoding transliteration and encoding hieroglyphs.

1. Unicode for transliteration.
In the 1980s and 1990s, Egyptologists used a variety of customized fonts for representing transliteration, with often incompatible encoding. This entailed a lot of work for the editors of journals. Worse, texts with different encodings cannot be easily searched; as modern computers allow easy and fast searches

on documents, this is a waste of resources. Many characters needed for transliteration, like ḫ or ẖ, were already in Unicode, as they are also used outside of Egyptology. Thanks mainly to the work of M. Everson, codes were given to aleph and ayin in Unicode 5.1. Some open-source fonts, like "Deja Vu Sans" or "Gentium," now include these characters (Unicode Latin D).

A number of different solutions have been officially accepted for encoding the "Egyptological yod," basically with an "i" and three possible accents (Unicode FAQ).

Sign	Name	Code
Ꜣ	Latin capital letter Egyptological aleph	U+A722
ꜣ	Latin small letter Egyptological aleph	U+A723
Ꜥ	Latin capital letter Egyptological ayin	U+A724
ꜥ	Latin small letter Egyptological ayin	U+A725

Although the Unicode standard defines uppercase and lowercase versions of aleph and ayin, this is a rather artificial creation, not corresponding to Egyptological uses. A font having the exact same glyph for the uppercase and lowercase variants would thus be faithful to actual practices.

Note that most publishing houses have developed their own private conventions regarding the use of Unicode, at a time when there were no codes for "aleph" and "ayin." To reduce the problems in the future, and in particular to ensure a unity of encoding, it would be preferable to drop specific conventions and to adopt the standard Unicode encoding as soon as possible.

2. Unicode for hieroglyphs.
The Unicode encoding of hieroglyphs started as early as 1994, and a final version has been included in Unicode 5.2 (2009) thanks to the work of B. Richmond and M. Everson (Everson 1999; Everson and Richmond 2007).

The encoding was proposed after an agreement by a number of scholars and a presentation to the "Informatique et égyptologie" working group in Oxford in 2006. As the possible content of an extensive list was still felt to be an open problem, the decision was taken to limit the encoding to the original Gardiner fonts, plus their various extensions until 1953, and a handful of other signs.

The current coverage of Unicode is not sufficient for monumental texts, especially ones from the Late Period, but is already usable for encoding hieratic texts, as the Gardiner sign list is more or less sufficient for this purpose. In any case, the current standard will ease the use of hieroglyphic fonts with a common encoding and also make parts of the encoded hieroglyphic texts searchable without needing specific tools. Texts encoded using the Manuel de Codage need specific software to be manipulated. Unicode-encoded texts, can (within some limitations) be used with commercial software such as MS/Word or OpenOffice. As an example, search results from the Ramsès project (Winand et al. fc.) can be exported to MS/Excel files to create arbitrary statistics. MS/Excel does not understand MdC encoding; by using Unicode, it was possible to include a rough yet readable approximation of the word spellings.

One should note that by definition Unicode does not include any mechanism for glyph positioning.

Sign Positioning

To encode hieroglyphic texts fully, the layout of the signs must also be represented. Most current encodings are based on the Manuel de Codage. The original Manuel uses the symbols "*" to indicate that two signs are on the same line, "-" to separate blocks, and ":" to stack two lines. Parenthesis allow complex quadrants with sub-groups. Thus ▱ can be written "Q3*X1:N1".

The original version of the Manuel lacked a number of possibilities. It had a very rough system for indicating damaged zones, nothing to indicate lines or columns, and did not allow complex grouping like ◿. Various software using the Manuel have provided proprietary extensions to deal with the problem. Thus, a number of incompatible dialects of the Manuel de Codage are now in use (in most cases, conversions are however possible).

Most software use very font-dependent encoding for representing groups like ◿ . In reaction to this, M.-J. Nederhof (2013) developed the Revised Encoding Scheme (RES): this includes new operators, like "insert," which states that a group should be "inserted" at a precise position in the square occupied by a sign. For instance, "insert[te](G39,N5)" inserts the N5 sign in the "top end" corner of sign G39, yielding ◿. The advantage is that the encoding is not dependent on the particular fonts. It requires more sophisticated algorithms than those used by the MdC systems, but this is not a real problem. The added complexity can be hidden to the scholar, and it is therefore likely that at least ideas from RES will find their way in future systems.

The JSesh software, developed by the present author (Rosmorduc 2014), uses a slightly extended version of the Manuel de Codage and supports a "ligature" system inspired by MacScribe. It is less powerful than the one proposed for RES, but still allows for groups like ▦ (encoded F20&&&[xAst:xAst: xAst]) to be automatically built, with two operators to combine a group of signs (here ▦) "before" or "after" a sign (here ◿).

In a rather different direction, G. Lapp created two editors, which combine drawing facilities with hieroglyphic editing, VisualGlyph and now iglyph (Subotic and Lapp 2008). This software is usable for near-facsimile drawing.

Other Systems

Other encoding systems have been proposed, partly inspired by the encoding used by Assyriologists. W. Schenkel (1983) uses an encoding, which enriches the transliteration

with information about sign groups and determinatives. The recent publication of the *Book of Caverns* by D. Werning (2011) has hieroglyphic determinatives as exponents at the end transliterations of Egyptian words: this shows that a mixture of transliteration and a few hieroglyphs can be relatively readable, much more than plain transliteration, and might be a viable alternative to full hieroglyphic transcripts for databases in some cases. These enriched transliterations can now be created fully with Unicode fonts, without specific software.

Text Databases and Dictionaries

Computer text databases have been developed since the late 1960s. The pioneer was W. Schenkel with the MAAT (Maschinelle Analyse Altägyptischer Texte) project in 1967. A number of databases have been created since, often by small teams or individuals.

Although the Manuel de Codage was written with text databases in mind (van den Berg 1988; Hallof and van den Berg 1992), many existing corpora do not include hieroglyphic spelling, the reason being that hieroglyphic encoding is very time-consuming.

The largest current project is probably the *Thesaurus Linguae Aegyptiae* (TLA; Grunert and Hafemann 1999), which includes a large number of lemmatized texts, side by side with the original digitized *Wörterbuch* files (DZA = Digitales Zettelarchiv). The TLA provides online access to the database, for simple or combined searches (two words in the same context), and a selection of tools for statistical analysis. Among the basic principles behind the TLA concept, one can note the decision to be as theory-neutral as possible in the mark-up tag set, especially regarding the verbal system (Hafemann 2003; Seidlmayer 2003). The database basically records the morphological features of the verbs, thus avoiding the need to interpret the forms. Although the TLA did not at first record the hieroglyphic spellings, it started to add spelling information in 2009 for lexical entries (Hafemann and Dils 2011; Dils and Feder 2013), and work is being done to add spelling information to the texts.

The Ramsès project (Rosmorduc et al. 2008; Winand et al. 2008) is a fully lemmatized database of Late Egyptian. Word analysis includes references to lemma, part-of-speech tagging, and hieroglyphic spelling. The texts are annotated with rich metadata, about both the text content (genre, for instance) and material support (type of support, writing system, date). The presence of metadata in large corpora is very important as it allows to build sub-corpora, and possibly to cross-query results. Syntactically analyzed corpora (also called "tree-banks") are not available yet for Egyptian. One of the goals of the Ramsès project is to create such a resource. Tools for creating such corpora have been created and tried out.

A number of projects deal with Demotic (Bresciani et al. 2002; DWL; Hoffmann 2005); the TLA includes Demotic documents, and, last but not least, the Chicago Demotic Dictionary (CDD) provides a searchable dictionary with spellings.

Currently, most institutional databases can be searched online. Their content is usually not openly available in raw form. Statistical searches and the like are only possible through the interface of their web sites, or in collaboration with the hosting institution.

A large number of databases have been produced by groups or individuals, sometimes outside academic circles. Among these, one can cite the Projet Rosette (Euverte 2008), which is an interesting example of web environment with a mainly pedagogical purpose. AELalign by M.-J. Nederhof (2008), a computational linguistic researcher, is an elaborate system to share annotated texts, which has been used by the Ancient Egyptian Language discussion list (AEL); his software is open-source and freely available, and includes a number of texts. In a less elaborate way, the JSesh hieroglyphic editor comes with around 150 raw texts in Manuel de Codage format.

Natural Language Processing

Natural Language Processing (NLP), as such, usually involves automated procedures applied

to a text. Although a number of works of research have been applied to Egyptian, very few of them have resulted in a usable system, the rest being model implementations. The current trend in NLP, since the late 1990s, has been to use statistical models, and the available text databases for Egyptian now have a reasonable size for using these. A number of current projects involve NLP components, in more or less ambitious forms. NLP is inherently error-prone. However, human-made annotations, even done by experts, also contain errors. Automated processing is often reliable enough to provide a first draft, which can then be proofread by humans. The actual details depend on the complexity of the task at hand.

1. Automated transliteration.

Automated transliteration was the subject of Sophie Billet's PhD Thesis (Billet 1995), using "intelligent agents," an artificial intelligence approach. A less sophisticated system, based on weighted rewriting rules and finite-state transducers, was created by Rosmorduc (2008), with around 80% correct transliterations if the original is in hieratic. Nederhof (2008) proposed a system to "align" a hieroglyphic transcription and an existing transliteration, which can be used to automatically produce an interlinear edition of a text, for instance. In this case, which is simpler than direct transliteration, more than 98% success was obtained on the test corpus.

2. Part-of-speech taggers and syntactic analysis.

Ad-hoc attempts at automated part-of-speech tagging were already made in the 1980s (Winand 1988). Automated part-of-speech taggers for modern languages can obtain more than 95% accuracy nowadays, and most tagged corpora are built by automated tagging followed by a human proofreading. Typically, those taggers use statistical algorithms and learn from an already tagged corpus (this is categorized as "supervised training"). These supervised algorithms need corpora of a few tens of thousands of words to be trained. Although no experimental result has been

published yet, the size of the existing computer corpora in Egyptology is now sufficient.

Automated parsing of Egyptian is still in its infancy. One can expect that the current development of tagged corpora will provide the needed source material. A theoretical issue for automated parsing is the choice of the grammatical framework (however, this is even more a problem for syntactic tree-banks if human processing is considered).

Statistical Analysis

Large text databases naturally lead to statistical processing. A number of tools are available for this kind of work, either general tools like Gnu R, or specialized ones like the Textométrie project (Heiden 2010). Statistical analysis results are often interesting starting points for reflection, but they must be handled with care. A good understanding of the biases of the statistical method used, and of the importance of the size of the sample, are needed. A collaboration with an actual statistician is probably a good idea. This is also true the other way around: the statistician can point out statistically significant phenomena, but Egyptological knowledge will be needed to distinguish between what is really relevant and what is trivial.

A number of attempts have been made in particular for text classification. Classically, one distinguishes unsupervised classification, where the classes emerge from the system, from supervised classification, where the system "trains" on a set of already classified texts and "learns" to classify new texts.

A simple statistical measure, called S*, is computed from the size of the text and the number of distinct words in it: this has been used in attempts to measure the richness of vocabulary (Lepper 2013, with a good bibliography on lexicostatistics in Egyptology), further interpreted as an indication of textual genre. This measure, by itself, is probably somewhat crude and does not allow to distinguish between factors such as language change, author vocabulary, text genre, or text register.

Statistics on text vocabulary have also been used as a strategy for dating (Schweitzer 2013). The idea is that, in theory, statistics on the whole vocabulary of a text should provide better criteria for dating than the study of individual words. Schweitzer's work uses the vector space model of texts, and hierarchical clustering, an unsupervised method, which builds trees of texts. Depending on how many words they share or not share, texts are assigned a distance with respect to one another. The closest texts are grouped together: a family tree of texts thus emerges, which is then used to distinguish classes. Given that the clustering results seem to be independent of text genre, Schweitzer proposed that it can be at least partly interpreted as reflecting diachronic features.

An experiment in text classification using a number of widespread machine learning methods is described in Gohy et al. (2013). This is a supervised experiment in which the various text genres are predefined. It contains a detailed analysis of the results, comparing the results for the methods according to the various text genres.

Optical Character Recognition

Automated recognition of hieroglyphic texts from an original drawing or photograph is still in its infancy, as a rather exotic use of Optical Character Recognition (OCR) in general. However, the progress in OCR in recent years, with advances in the combination of image processing, statistical methods, and linguistic information, has lead to interesting results.

Franken and van Gemert (2013) experimented on photographs of the pyramid texts of Unas as corpus and succeeded in matching the signs with those found in the hieroglyphic font of JSesh, using a lexicon to arbitrate between concurrent solutions. The corpus used is of course a rather favorable case for OCR, as the signs in the pyramid texts of Unas are very clear and well-formed, but this represents an interesting starting point. Even with an imperfect OCR, a number of applications can be considered. For instance, scans of old Egyptological journals could be improved for searching individual signs. Another technique, called word spotting (Rath and Manmatha 2003), which is purely image-based, has been used with good results for text exploration in various kinds of documents, for instance, Latin medieval manuscripts. In OCR, a picture of a text is analyzed in order to find the original text. Word spotting does something simpler for a computer: it searches in a picture database (for instance, the *Book of the Dead* papyri) all places where a given word can be found. However, to start the search, one needs a picture of the word in question; this method is therefore mainly suited for cross-references and indexes. Recently, an OCR System, created by Nederhof, has been included in his PhilologEg open-source system.

Bibliographic Notes

The special issue of the Atala journal on Natural Language Processing and ancient languages (Denooz and Rosmorduc 2009) is a good starting point for the uses of NLP in the humanities. Among the non-Egyptological articles in this issue, Haug et al. (2009) provides an interesting example of linguistic database conception. The various proceedings of the computer working group (Strudwick 2008; Polis and Winand 2013) contain a number of interesting articles illustrating the various domains of application of computational linguistics in Egyptology, notably text encoding, text databases, automated transliteration, text alignment, and statistical uses of databases. Jurafsky and Martin (2008) gives an up-to-date, but very technical, description of NLP, and is one of the standard manuals. For statistical approaches Manning and Schütze (1999) is a classical textbook with NLP orientation; Baayen (2008) gives a practical approach, orientated toward a linguistic audience. For pure statistical processing, Gries (2013) provides an introduction specialized for linguists.

References

Baayen, R. H.
  2008   *Analyzing linguistic data: A practical introduction to statistics using* R. Cambridge: Cambridge University Press.

Baines, John, and Catherine Griffin
  1988   Automated typesetting of the Gardiner hieroglyphic font. *Informatique et Egyptologie* 4, pp. 85-95.

van den Berg, Hans
  1988   Textual analysis on computer: Some additions to the encoding system of the manual. *Informatique et Egyptologie* 4, pp. 32-43.

Billet, Sophie
  1995   *Apports à l'acquisition interactive de connaissances contextuelles*. PhD dissertation: Université Montpellier II.

Bresciani, Edda, Andrea Menchetti, Paolo Bozzi, Alfredo Ruffolo, Giuseppe Eisinberg, and Giuseppe Fedele
  2002   *Computational philology system for Demotic texts*. (XIVth meeting of the Computer Working Group of the International Association of Egyptologists. Available on CDROM only).

Buurman, Jan, Nicolas Grimal, Michael Hainsworth, Jochen Hallof, and Dirk van der Plas
  1988   *Inventaire des signes hiéroglyphiques en vue de leur saisie informatique*. Mémoires de l'Académie des Inscriptions et Belles Lettres: Nouvelle Série 8. Paris: De Boccard.

Cauville Sylvie
  1997   *Le temple de Dendara: Les chapelles osiriennes*. Cairo: Institut français d'archéologie orientale.

Denooz, Joseph, and Serge Rosmorduc
  2009   Preface: Traitement automatique des langues et langues anciennes. *Traitement Automatique des Langues* 50(2). (Internet resource: http://www.atala.org/IMG/pdf/TAL-2009-50-2-00-Preface-en.pdf.)

Dils, Peter, and Frank Feder
  2013   The Thesaurus Linguae Aegyptiae: Reviews and perspectives. In *Texts, languages & information technology in Egyptology*, ed. Stéphane Polis and Jean Winand, pp. 103-110. Liège: Presses Universitaires de Liège.

Euverte, Vincent
  2008   The Rosette Project: Computer assistance for the student, the epigraphist and the philologist. In *Information technology and Egyptology in 2008: Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists (Vienna, 8-11 July 2008)*, ed. Nigel Strudwick, pp. 71-92. Piscataway, N.J.: Gorgias Press.

Everson, Michael
  1999   *Encoding Egyptian hieroglyphs in Plane 1 of the UCS*. (Internet resource: http://www.unicode.org/L2/L1999/N1944.pdf)

Everson, Michael, and Bob Richmond
  2007   *Proposal to encode Egyptian hieroglyphs in the SMP of the UCS*. Working Group Document ISO/IEC JTC1/SC2/WG2 N3237, International Organization for Standardization. (Internet resource: http://std.dkuug.dk/jtc1/sc2/wg/docs/n3237.pdf.)

Franken, Morris, and Jan van Gemert
  2013   Automatic Egyptian hieroglyph recognition by retrieving images as texts. In *Proceedings of 21st ACM international conference on Multimedia* (ACM MM 2013), pp. 765-768. New York: ACM.

Gohy, Stéphanie, Benjamin Martin Leon, and Stéphane Polis
  2013   Automated text categorization in a dead language, the detection of genres in Late Egyptian. In *Texts, languages & information technology in Egyptology*, ed. Stéphane Polis and Jean Winand, pp. 61-74. Liège: Presses Universitaires de Liège.

Grunert, Stefan, and Ingelore Hafemann
1999   *Textcorpus und Wörterbuch: Aspekte zur ägyptischen Lexikographie.* Probleme der Ägyptologie 14. Leiden, Boston, and Cologne: Brill.

Gries, Stefan
2013   *Statistics for linguistics with R: A practical introduction.* Mouton Textbook. Berlin: De Gruyter.

Hafemann, Ingelore (ed.)
2003   *Wege zu einem digitalen Corpus ägyptischer Texte: Akten der Tagung "Datenbanken im Verbund"* (Berlin, 30. September - 2. Oktober 1999). Thesaurus Linguae Aegyptiae 2. Berlin: Achet Verlag.

Hafemann, Ingelore, and Peter Dils
2011   Der Thesaurus Linguae Aegyptiae – Konzepte und Perspektiven. In *Perspektiven einer corpusbasierten historischen Linguistik und Philologie: Internationale Tagung des Akademienvorhabens "Altägyptisches Wörterbuch" an der BBAW, 12.-13. Dezember 2011,* Thesaurus Linguae Aegyptiae 4, ed. Ingelore Hafemann, pp. 127-144. Berlin: Achet Verlag.

Hallof, Jochen
1988   Probleme bei der Codierung von ägyptischen Schriftzeichen. *Informatique et égyptologie* 5, pp.11-18.

Hallof, Jochen, and Hans van den Berg
1992   Thot: Zum Konzept eines Programs zur Analyse ägyptischer Texte. *Informatique et égyptologie* 8, pp. 42-46.

Haug, Dag, Marius Jøhndal, Hanne Eckhoff, Eirik Welo, Mari Hertzenberg, and Angelika Müth
2009   Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages. *TAL* 50, no. 2, pp. 17-45.

Heiden, Serge
2010   The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme. In *Proceedings of the 24th Pacific Asia Conference On Language, Information and Computation*, Sendai, Japan, ed. Ryo Otoguro et al., pp. 389-398. Tohoku, Japan: PACLIC 24.

Hoffmann, Friedhelm
2005   Ein demotistisches EDV-Werkzeug: Die Demotische Wortliste (DWL). In *Actes du ixe congrès international des études démotiques, Paris, 31 août-3 septembre 2005*, Bibliothèque d'étude 147, ed. Ghislaine Widmer and Didier Devauchelle, pp. 145-156. Cairo: Institut français d'archéologie Orientale.

Janssen, Josef
1972   Les listes de signes hiéroglyphiques. In *Textes et langages de l'Égypte pharaonique*, ed. Serge Sauneron, pp. 57-66. Cairo: Institut français d'archéologie orientale.

Jurafsky, Daniel, and James Martin
2008   *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 2nd edition. Upper Saddle River, N.J.: Pearson Education.

Lepper, Verena.
2013   Ancient Egyptian literature: Genre and style. In *Ancient Egyptian literature: Theory and practice,* ed. Roland Enmarch and Verena Lepper, pp. 211-225. Oxford: Oxford University Press.

Manning, Christopher, and Hinrich Schütze
1999   *Foundations of statistical natural language processing.* Cambridge, Mass.: MIT Press.

Meeks, Dimitry
1995   Le logiciel S.E.C.H.A.T. et la diffusion de l'information en égyptologie. In *Aplicaciones informaticas en arqueologia. Teorias y sistemas*, Vol. I, pp. 308-311. Bilbao: Denboraren Argia.

Nederhof, Mark-Jan
2008   Automatic alignment of hieroglyphs and transliteration. In *Information technology and Egyptology in 2008: Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists*, Vienna, 8-11 July, 2008, ed. Nigel Strudwick, pp. 71-92. Piscataway, N.J.: Gorgias Press.

2013   The Manuel de Codage encoding of hieroglyphs impedes development of corpora. In *Texts, languages & information technology in Egyptology*, ed. Stéphane Polis and Jean Winand, pp. 103-110. Liège: Presses Universitaires de Liège.

Piotrowski, Michael
2012   *Natural language processing for historical texts*. San Rafael, Calif.: Morgan & Claypool.

Polis, Stéphane, and Serge Rosmorduc
2013   Réviser le codage de l'égyptien ancien: Vers un répertoire partagé des signes hiéroglyphiques. *Document Numérique* 16(3), pp. 45-69.

Polis, Stéphane, and Jean Winand (eds.)
2013   *Texts, languages & information technology in Egyptology*. Liège: Presses Universitaires de Liège.

Rath, Toni, and R. Manmatha
2003   Features for word spotting in historical manuscripts. In *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR), Edinburgh, Scotland, August 3-6, 2003*, Vol. 1, pp. 218-222. Los Alamitos, Calif.: IEEE Computer Society.

Rosmorduc, Serge
2002   Le codage informatique des langues anciennes, le cas des hiéroglyphes égyptiens. *Documents Numériques* 6(3-4), pp. 211-225.
2008   Automated transliteration of Egyptian hieroglyphs. In *Information technology and Egyptology in 2008: Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists (Vienna, 8-11 July 2008)*, ed. Nigel Strudwick, pp. 167-183. Piscataway, N.J.: Gorgias Press.
2014   *JSesh Documentation*. (Internet resource: http://jseshdoc.qenherkhopeshef.org.)

Rosmorduc, Serge, Stéphane Polis, and Jean Winand
2008   Ramses: A new research tool in philology and linguistics. In *Information technology and Egyptology in 2008: Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists (Vienna, 8-11 July 2008)*, ed. Nigel Strudwick, pp. 155-166. Piscataway, N.J.: Gorgias Press.

Schenkel, Wolfgang
1983   *Aus der Arbeit an einer Konkordanz zu den altägyptischen Sargtexten*: Teil 1: Zur Transkription des Hieroglyphisch-Ägyptischen. Göttinger Orientforschung IV: Ägypten 12. Wiesbaden: Harrassowitz.

Schweitzer, Simon
2013   Dating Egyptian literary texts: Lexical approaches. In *Dating Egyptian literary texts: Proceedings of the Conference Göttingen 9.-12.6.2010*, Lingua Aegyptia Studia Monographica 11, ed. Gerald Moers, Antonia Giewekemeyer and Kai Widmaier, pp. 177-190. Hamburg: Widmaier Verlag.

Seidlmayer, Stefan
2003   Textdatenbanken im Verbund – Konzepte und Perspektiven. In *Wege zu einem digitalen Corpus ägyptischer Texte: Akten der Tagung "Datenbanken im Verbund" (Berlin, 30. September - 2. Oktober 1999)*, Thesaurus Linguae Aegyptiae 2, ed. Ingelore Hafemann, pp. 207-235. Berlin: Achet.

Stief, Norbert
1988   Weitere Möglichkeiten bei der Hieroglyphenausgabe via Computer. *Informatique et Egyptologie* 5, pp. 46-51.

Strudwick, Nigel (ed.)
2008   *Information technology and Egyptology in 2008: Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists (Vienna, 8-11 July 2008)*, pp. 71-92. Piscataway, N.J.: Gorgias Press.

Subotic, Ivan, and Günther Lapp
2008   iGlyph: A hieroglyphic text processing program. (Internet resource: http://code.google.com/p/iglyph/)

Tiradritti, Francesco
    1988   Program for automatic analysis and printing of hieroglyphic texts (suggestions for a Coding). *Informatique et Egyptologie* 5, pp. 52-65.

Winand, Jean
    1988   Lemmatisation et levée d'ambiguïtés automatique (II). *Informatique et Egyptologie* 5, pp. 76-92.

Winand, Jean, Stéphane Polis, and Serge Rosmorduc
    fc.   Ramses: An annotated corpus of Late Egyptian. In *Proceedings of the Xth IAE Congress (Rhodos, 2008)*, ed. Panagiotis Kousoulis. Leuven: Peeters.

Werning, Daniel
    2011   *Das Höhlenbuch: Textkritische Edition und Textgrammatik*. 2 vols. Göttinger Orientforschung IV: Ägypten 48. Wiesbaden: Harrassowitz.

External Links

AEL
    Ancient Egyptian Language discussion list. (Internet resource: http://www.rostau.org.uk/AEgyptian-L/. Accession date: November 2014.)

Chicago Demotic Dictionary
    The Chicago Demotic Dictionary (CDD). (Internet resource: http://oi.uchicago.edu/research/projects/chicago-demotic-dictionary-cdd-0. Accession date: December 2014.)

DWL
    Demotische Wortliste. (Internet resource: http://www.dwl.aegyptologie.lmu.de/.   Accession date: November 2014.)

DZA
    Digitalisierung des Zettelarchivs der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) zum Wörterbuch der ägyptischen Sprache mit Indizierung und Sicherheitsverfilmung (Altägyptisches Wörterbuch). (Internet resource: http://www.kulturerbe-digital.de/en/projekte/9_38_363239.php. Accession date: November 2014.)

EPIDOC
    EpiDoc: Epigraphic documents in TEI XML. (Internet resource: http://sourceforge.net/p/epidoc/wiki/Home/ Accession date: November 2014.)

GNU R
    The R Project for statistical computing. (Internet resource: http://www.r-project.org. Accession date: November 2014.)

iGlyph
    iGlyph. A hieroglyphic text processing program. (Internet resource: http://code.google.com/p/iglyph/. Accession date: November 2014.)

JSesh
    JSesh, an open source hieroglyphic editor. (Internet resource: http://jsesh.qenherkhopeshef.org/. Accession date: November 2014.)

PhilologEg
    PhilologEg. (Internet resource: http://mjn.host.cs.st-andrews.ac.uk/egyptian/align/. Accession date: November 2014.)

TEI
    TEI: Text Encoding Iniative. (Internet resource: http://www.tei-c.org/. Accession date: November 2014.)

Thesaurus Linguae Aegyptiae
    Arbeitsstelle Altägyptisches Wörterbuch. Berlin-Brandenburgische Akademie der Wissenschaften.
    (Internet resource: http://aaew.bbaw.de/tla/. Accession date: November 2014.)

Unicode Latin D
    Unicode Latin Extended D. (Internet resource http://www.unicode.org/charts/PDF/UA720.pdf.
    Accession date: December 2014.)

Unicode FAQ
    Unicode Frequently Asked Questions, about "combining marks". (Internet resource
    http://www.unicode.org/faq/char_combmark.html#20. Accession date: December 2014.)