12-22-2010

# Linguistic Informatics: A Humanistic Endeavor

Jan H. Kroeze
*University of South Africa*, jan.kroeze@gmail.com

# Linguistic Informatics: A Humanistic Endeavor

Jan H. Kroeze
University of South Africa, South Africa

**Abstract**
This paper discusses the place of research on the use of information systems technologies to store and explore linguistic data. As such, it may be regarded as small contribution to the philosophy of science. After a section on the nature of informatics in general, some general principles will be applied to computer-assisted linguistic studies, and a number of examples will be provided to illustrate the concept of linguistic informatics.

**Keywords:** Information Systems, Linguistic Data

**Permanent URL:** http://sprouts.aisnet.org/7-19

**Reference:** Kroeze, J.H. (2007). "Linguistic Informatics: A Humanistic Endeavor," . *Sprouts: Working Papers on Information Systems*, 7(19). http://sprouts.aisnet.org/7-19

## Introduction

Informatics is an interdisciplinary science that explores the application and effect of information technology in business, organisations and society. It is regarded as a social science with its main focus on the human aspects in the symbiotic relationship of computers and people. Therefore, it uses methodologies and research paradigms that are typical of the humanities, such as qualitative research and anti-positivistic points of departure.

It is known phenomenon that information technology changes the working culture and structure of organisations (see Du Plooy, 1998: 12-23). A similar phenomenon can be observed in other humanistic computing areas. One of these areas is computational linguistics, which studies the use of computer technology to enhance the study of language. In computational linguistics the researcher may either focus on the algorithmic simulation of language rules and production, or on the efficient storage and use of linguistic data which has already been analysed and captured. First-mentioned will typically be studied by computer scientists, while last mentioned will typically be investigated by informaticians, since it focuses on the data storage by means of databases, as well as the exploration of that data by means of data mining and data warehousing ventures.

The study of databases forms part of the discipline of informatics or information systems (IS). According to Vessey et al. (2002: 167) database management is one of the topics that are “at the heart of the IS discipline in that they are central to IS curricula and therefore to IS careers”. The creation of knowledge databases and the exploration of these electronic repositories are thus part and parcel of information systems research, even if the encoded data come from other disciplines. “The *power* and not the weakness of IS research models is precisely that they situate IS constructs within constructs that other disciplines study” (Agarwal & Lucas, 2005: 390). Since information systems itself is an interdisciplinary science it should not only aim to add value to other disciplines, but also borrow from other contributing ICT[1] disciplines in order to strengthen their allegiances.

This paper discusses the place of research on the use of information systems technologies to store and explore linguistic data. As such, it may be regarded as small contribution to the philosophy of science. After a section on the nature of informatics in general, some general principles will be applied to computer-assisted linguistic studies, and a number of examples will be provided to illustrate the concept of linguistic informatics.

## Informatics as a humanistic research field

Although it is tempting to view information technology as “the epitome of rational expression”, positivist research is only apt to study the harder engineering and algorithmic issues related to it. These issues are indeed studied by the “harder” natural sciences, such as computer science and computer engineering. Socially constructed issues such as information systems software should, however, rather be studied from

[1] Information and communication technology.

the angle of humanistic enquiry. This is especially true of data mining ventures because new knowledge is not simply discovered, but created (cf. Du Plooy 1998: 54, 59).

Cilliers (2005) pleads for alternative forms of knowledge regarding complex systems which are modest and provisional, acknowledging that our understanding is limited and changing. Restricting humanistic research to scientific objectivism and crude positivism would be unethical and dishonest because it would pretend that this knowledge is final and beyond all dispute. Modest claims about knowledge, however, invite knowledge workers to persevere in an ongoing search for meaning and generation of understanding.

In informatics there has indeed been a growing acceptance that positivist research is not the only valid scientific paradigm that could be used to produce good research. Avgerou (2005:105), for example, argues for critical research using interpretive methods in information systems to complement empirical and formal cognitive methods. She regards critical research as a process that aims to make sense of the investigated scenario, a radical procedure in which researchers' human capacities such as tacit knowledge and moral values are involved. "I see research as the art of putting together research questions with a critical content, multiple theories and epistemological awareness to develop claims of truth. This art cannot place confidence for producing valid knowledge on adhering to a testable theory or research practice" (ibid.: 108). Although the knowledge claims contributed by interpretive case studies should be regarded as soft facts, they are still valid and should be generalised in clear formulations aimed at identified target audiences (Barret & Walsham, 2004: 298, 310).

Bondarouk & Ruël (2004) argues for the use of discourse analysis to do research on information systems documents. Discourse analysis is another non-objectivist hermeneutical method. It is essentially interpretive and constructivist. It tries to "give a meaning to a text within a framework of the interpreter's experience, knowledge, time, epoch, culture, and history". It believes that understanding is an open, continuous process and that there is no final, authoritative interpretation.

Other anti-positivist approaches in informatics, which will not be discussed here in detail, are (cf. Carlsson, 2003; Du Plooy, 1998: 53-68):

- grounded theory (the researcher derives theory by means of qualitative data analysis)
- ethnography (the researcher participates in activities of the organisation that is studied)
- action research (the researcher collaborates with members of the organisation to experiment with possible solutions for a problem)
- structuration theory (the researcher regards human agency and social structure as an inseparable duality)
- critical realism and adaptive theory (the researcher attempts to combine and synthesise positivism and interpretivism)
- actor network theory (the researcher studies the technical and social aspects of IT as a unity because values are believed to be built into software)[2]

[2] Like adaptive theory which is epistemologically neither positivist nor interpretivist (Carlsson, 2003), actor network theory (ANT) is positioned between deterministic and constructivist theories (Cordella & Shaikh,

## Linguistic informatics as a humanities discipline

From the discussion above it should already be clear that is has become acceptable to use softer, interpretive methods in informatics research. The representation of data, including linguistic data, is one of the basic ventures of humanities computing (Neyt, 2006: 2-5). It may be a tool that could introduce a “softer” view and use of computers that would be more applicable in the humanities than the “harder” approaches that are typical of the natural sciences.

Ramsay (2003) proposes an “algorithmic criticism”, which rejects the use of computers only to empirically confirm or reject hypotheses, because it constrains meaning. He suggests that computing humanists should rather use software to discover a multiplicity of meanings in literary sources. Such an approach will deepen the subjectivity that is essential for the creation of critical insight. In order to reach this goal the researcher must be able to perform a playful exploration of the text that reveals exciting new patterns built on re-orderings of marked-up text. “It is, of course, possible to go to a literary text armed with a hypothesis, but we do better to go to it with a hunch borne of our collective musings – a sneaking suspicion that looking at it *this* way will turn up something interesting. Or better still, we could go to it with a machine that is ready to reorganize that text in a thousand different ways instantly” (Ramsay, 2003: 171). Researchers and software creators should therefore work towards alternatives for the traditional, statistics-based “forensic semiotics” in the processing of texts in order to change the computer into a tool that support interpretive processes: “[R]ather than to extol the computer as a scientific tool that can supposedly help prove particular facts about a text, we would do better to focus on its ability to help read, explore, experiment, and play with a text” (Sinclair, 2003: 176).

## Some examples of linguistic informatics

Any software tool that allows the researcher to adopt a more holistic approach may be regarded as a linguistic information system. This definition is in line with an externalist view[3] of good science which approves the incorporation of insights from other disciplines (Dennis et al., 2006: 7-8). One should, of course, explore the possibility of three- or multi-dimensional data layers in the software to render linguistic analyses, because “[i]t is our conjecture that linguistic meaning is intrinsically and irreducibly very high dimensional” (Landauer et al., 2004: 5214).

A simple representation of the phonetics, morphology, syntax and semantics of a specific text using the web language *html* may be regarded as a very simple linguistic information system, since it uses mark-up tags to format and organise this data. Kroeze (2002), for example, used *html* to create a web-based program that

2003). It studies the reciprocal influence of technology and society, the interaction between the human and non-human actors that constitute a network. Reality is believed to come into existence through this interplay.

[3] An internalist view, on the other hand, argues “that a core set of knowledge and shared scientific paradigms generated internal [*sic*] to the discipline are hallmarks of mature science, and thus diversity is to be avoided” (Dennis et al., 2006: 7).

shows analyses of the Hebrew text of Jonah in an interlinear fashion as a series of tables. This project illustrates how modern information technology can be used to publish the results of old and new analytical techniques used in the analysis of an ancient language.

Computer-assisted reading and text synthesis are other ways of making the use of computers more acceptable to humanistic scholars. Computer-assisted reading could be regarded as a more advanced representation than an interlinear approach, which is however still text-based. Computer-assisted text synthesis could be used in the report function of a visualisation tool to produce suitable outputs of research results, to reproduce the original text (without mark-up and analysis), or to create an amended text according to the end-users' requirements (cf. Sinclair, 2003: 178-180).

Although these technologies allow the dissemination and retrieval of linguistic information, it does not facilitate more advanced capabilities to manipulate and investigate the data. Text mining is another way of dealing with textual information. Text-mining is still a new field in IT (especially in South Africa), but indications are that it will become a very important area, because the majority of business intelligence is stored in unformatted text-format. Techniques that will enable companies to mine for undiscovered valuable nuggets of information may become a new weapon to gain competitive advantage. Text-mining will also be used by information scientists and researchers in various fields to acquire new knowledge, such as the identification of unknown trends and associations between concepts or entities. However, text mining often uses computer science algorithms, and therefore represents the harder approach towards computational linguistics. However, it does prove that more advanced textual exploration is possible; it also suggests that the solution is to be found in data mining and its related technologies.

Closely related to data-mining and text-mining is data warehousing, which the author believes could offer an efficient solution for the effective processing of linguistic data in order to facilitate grammatical studies. Programming concepts that are typical of data warehousing may be adapted and used to store and explore this type of data sets. A three dimensional data cube, could, for example be used to store unlimited layers of linguistic data per clause, using the clauses as rows, the words or phrases as columns, and the third dimension for the various linguistic analyses related to these elements. Such a data bank may then be sliced and diced to reveal required combinations of linguistic modules. XML is a mark-up language that allows the programmer to develop his/her own hierarchical set of tags, thus facilitating the creation of a multi-dimensional database which may be sent as a flat file over the internet. The patterns that the researcher wants to unveil are covertly embedded within other visible patterns, i.e. the overt patterns specified by the schema. The XML schema defines the structure and content of the databank containing the XML mark-up tags (Clark et al., 2003).[4] "TEI tags are not merely structural delineations, but patterns of potential meaning woven through a text by a human interpreter" (Ramsay, 2003: 171).

[4] An XML schema is preferred above a DTD since it is more advanced and "more closely maps to database terminology and features" allowing the definition of variable types and valid values for the elements (Rob & Coronel, 2007: 579).

A linguistic database can therefore be stored and transported in XML format efficiently. A program can be created in a third generation computer language to extract the data by using a set of string processing functions. After the data has been converted into, for example, a three dimensional array, more advanced processing becomes available, such as slicing and dicing, rotation and drilling down. Routines may also be created to fulfil unique requirements of the researcher, such as to identify unique semantic role frameworks in a text, or to study the mappings of semantic roles onto syntactic functions. This implies that linguistic data is transformed into information, which can then be interpreted by the researcher to create knowledge.

According to Sinclair (2003: 182) computer-assisted play is a suitable method for humanistic computing of literary text. Such a playful exploration stimulates creativity which is necessary to improve linguists' understanding of language as a complex social system. Although "our understanding of complex systems cannot be reduced to calculation", the creative activity, on the other hand, should be a "careful and responsible development of the imagination" (Cilliers, 2005: 264). A visualisation tool, which allows the researcher to visually explore a set of linguistic data, could probably be used to stimulate these imaginative processes in such a responsible way.

## Conclusion

The study of linguistic information systems, linguistic informatics, may be regarded as a humanistic research endeavour. It investigates the possibilities to facilitate and enhance advanced processing and analysis of linguistic data in order to find hidden patterns which are difficult for humans to uncover. This information may then again be used to enrich the current knowledge of linguistics and languages. Electronic tools that facilitate linguistic data exploration will enable researchers to do more efficient and in-depth research on these phenomena, because it will facilitate and speed-up the gathering of extensive, relevant data to test hypotheses, or even to prompt new hypotheses.

## Bibliography

AGARWAL, R. & LUCAS, H.C. 2005. The information systems identity crisis: focusing on high-visibility and high-impact research. *MIS quaterly,* vol. 29, no. 3, pp. 381-398.

AVGEROU, C. 2005. Doing critical research in information systems: some further thoughts. *Info Systems J,* vol. 15, pp. 103-109.

BARRETT, M. & WALSHAM, G. 2004. Making contributions from interpretive case studies: examining processes of construction and use. In *Information systems research: relevant theory and informed practice (IFIP International Federation for Information Processing),* edited by Kaplan, B., Truex, D.P., Wastell, D., Wood-Haper, A.T. & DeGross, J.I., Part 3: Critical interpretive studies, Kluwer, pp. 293-312.

BONDAROUK, T. & RUËL, H. 2004. Discourse analysis: making complex methodology simple. In *Proceedings of the 12th European Conference on Information Systems (ECIS), June 14-16, 2004, Turku, Finland*, edited by Leino, T., Saarinen, T., & Klein S. [Online.] Available: http://www.csrc.lse.ac.uk/asp/aspecis/20040025.pdf [Cited 29 May 2006].

CARLSSON, 2003. Critical realism: a way forward in IS research. *Proceedings of the 11th European Conference on Information Systems, ECIS 2003, Naples, Italy 16-21 June 2003.* [Online.] Available: http://is2.lse.ac.uk/asp/aspecis/20030152.pdf [Cited 29 May 2006].

CILLIERS, P. 2005. Complexity, deconstruction and relativism. *Theory, culture and society,* vol. 22, no. 5, pp. 255-267.

CLARK, J., COWAN, J. & MAKOTO, M. 2003. RELAX NG compact syntax tutorial. Working draft 26 March 2003. [Online.] Available: http://www.relaxng.org/compact-tutorial-20030326.html [Cited 15 March 2006].

CORDELLA, A. & SHAIKH, M. 2003. Actor network theory and after: what's new for IS research? *Proceedings of the 11th European Conference on Information Systems, ECIS 2003, Naples, Italy 16-21 June 2003.* [Online.] Available: http://is2.lse.ac.uk/asp/aspecis/20030037.pdf [Cited 29 May 2006].

DENNIS, A.R., VALACICH, J.S., FULLER., M.A. & SCHNEIDER, C. 2006. Research standards for promotion and tenure in information systems. *MIS quaterly,* vol. 30, no. 1, pp. 1-12.

DU PLOOY, N.F. 1998. An analysis of the human environment for the adoption and use of information technology. Thesis (D.Com (Informatics)) – University of Pretoria.

KROEZE, J.H. 2002. Developing a multi-level analysis of Jonah using html. In *Bible and computer: the Stellenbosch AIBI-6 conference, Proceedings of the Association Internationale Bible et Informatique "From Alpha to Byte", University of Stellenbosch 17-21 July, 2000,* edited by J. Cook, Leiden: Brill, pp. 653-662.

LANDAUER, T.K., LAHAM, D. & DERR, M. 2004. From paragraph to graph: latent semantic analysis for information visualisation. *Proceedings of the National Academy of Science of the United States of America,* vol. 101, supplement 1, pp. 5214-5219.

NEYT, V. 2006. Fretful tags amid the verbiage: issues in the representation of modern manuscript material. *Literary and linguistic computing advance access,* pp. 1-13.

RAMSAY, S. 2003. Toward an algorithmic criticism. *Literary and linguistic computing,* vol. 18, no. 2, pp. 167-174. (Special section: reconceiving text analysis.)

ROB, P. & CORONEL, C. 2007. *Database systems: design, implementation, and management, 7th ed.* Boston, MA: Course Technology.

SINCLAIR, S. 2003. Computer-assisted reading: reconceiving text analysis. *Literary and linguistic computing,* vol. 18, no. 2, pp. 175-184.

VESSEY, I., RAMESH, V. & GLASS, R.L. 2002. Research in information systems: an empirical study of diversity in the discipline and its journals. *Journal of management information systems,* vol. 19, no. 2, pp. 129-174.

芽|Sprouts

Working Papers on Information Systems | ISSN 1535-6078

*Office*:
Sprouts
University of Amsterdam
Roetersstraat 11, Room E 2.74
1018 WB Amsterdam, Netherlands
Email: admin@sprouts.aisnet.org