LINGUISTIC INFORMATION – A HUMANISTIC ENDEAVOUR

by Jan H Kroeze

Informatics is an interdisciplinary science that explores the application and effect of IT in business, organisations and society. It is regarded as a social science that focuses on the human aspects in the symbiotic relationship between computers and people. Informatics uses methodologies and research paradigms typical of the humanities, such as qualitative research and anti-positivistic points of departure.

Information technology changes the working culture and structure of organisations (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 already analysed and captured.

Algorithmic simulation will typically be studied by computer scientists, while efficient data storage will typically be investigated by informaticians, since it focuses on 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 informatics or information systems (IS). According to Vessey et al. (2002: 167), database management is one of the topics "at the heart of the IS discipline in that it is central to IS curricula and therefore to IS careers."

The creation of knowledge databases and the exploration of these electronic repositories are part 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 is an interdisciplinary science, it should aim to add value to other disciplines and borrow from other contributing information and communication technology (ICT) disciplines to strengthen their allegiances.

Research was conducted to examine the place of research on the use of information systems technologies to store and explore linguistic data. It may be regarded as a small contribution to the philosophy of science.

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 studied by the "harder" natural sciences, such as computer science and computer engineering. Socially constructed issues such as information systems software should, however, be studied from the angle of humanistic enquiry. This is especially true of data-mining ventures, because new knowledge is not simply discovered, but created (Du Plooy 1998: 54, 59).

Cilliers (2005) pleads for alternative forms of knowledge regarding complex systems that 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 the generation of understanding.

In informatics there has been a growing acceptance that positivist research is not the only valid scientific paradigm that could produce good research. Avgerou (2005:105), for example, argues for critical research by 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.

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) argue for the use of discourse analysis to perform 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.

Linguistic informatics as a humanities discipline

It is 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 essential for the creation of critical insight.

Researchers and software developers should therefore work towards finding alternatives for the traditional, statistics-based "forensic semiotics" in the processing of texts to

change the computer into a tool that supports interpretive processes: "Rather 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 of good science that approves the incorporation of insights from other disciplines (Dennis et al., 2006: 7-8). 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, 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. 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).

One should, of course, explore the possibility of three-dimensional or multi-dimensional data layers in the software to render linguistic analyses, because "it is our conjecture that linguistic meaning is intrinsically and irreducibly very high dimensional" (Landauer et al., 2004: 5214).

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 to enable companies to mine for undiscovered valuable nuggets of information may become a new weapon to gain a competitive advantage.

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 to facilitate grammatical studies.

Programming concepts 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, 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.

The patterns 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 data bank containing the XML mark-up tags (Clark et al., 2003). 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). "TEI tags are not merely structural delineations, but patterns of potential meaning, woven through a text by a human interpreter" (Ramsay, 2003: 171).

According to Sinclair (2003: 182), computer-assisted play is a suitable method for the humanistic computing of literary texts. Such a playful exploration stimulates the creativity necessary to improve linguists' understanding of language as a complex social system.

Conclusion of the research

The study of linguistic information systems, linguistic informatics, may be regarded as a humanistic research endeavour. It investigates the possibilities of facilitating and enhancing the advanced processing and analysis of linguistic data in order to find hidden patterns that are difficult for humans to uncover.

This information may 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 they will facilitate and speed up the gathering of extensive, relevant data to test hypotheses, or even to prompt new hypotheses.

References

Agarwal, R. & Lucas, H.C. 2005. The information systems identity crisis: focusing on high-visibility and high-impact research. MIS quarterly, 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 quarterly, 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.

***Professor Jan H Kroeze** is associated with the University of Pretoria's Department of Informatics.*

jan.kroeze@up.ac.za