Zoubida Kedad Nadira Lammari
Elisabeth Métais Farid Meziane
Yacine Rezgui (Eds.)

Natural Language Processing and Information Systems

LNCS 4592

12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007
Paris, France, June 2007, Proceedings

Springer

Zoubida Kedad Nadira Lammari
Elisabeth Métais Farid Meziane
Yacine Rezgui (Eds.)

Natural Language Processing and Information Systems

12th International Conference on Applications
of Natural Language to Information Systems, NLDB 2007
Paris, France, June 27-29, 2007
Proceedings

Springer

Volume Editors

Zoubida Kedad
Laboratoire PRiSM, Université de Versailles, France
E-mail: Zoubida.kedad@prism.uvsq.fr

Nadira Lammari
Elisabeth Métais
Conservatoire National des Arts et Métiers (CNAM)
75141 Paris cedex 3, France
E-mail: {lammari, metais}@cnam.fr

Farid Meziane
University of Salford, Greater Manchester, UK
E-mail: f.meziane@salford.ac.uk

Yacine Rezgui
University of Salford, Informatics Research Institute
Greater Manchester, UK
E-mail: y.rezgui@salford.ac.uk

Library of Congress Control Number: 2007929429

CR Subject Classification (1998): H.2, H.3, I.2, F.3-4, H.4, C.2

LNCS Sublibrary: SL 3 – Information Systems and Application,
incl. Internet/Web and HCI

ISSN	0302-9743
ISBN-10	3-540-73350-7 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-73350-8 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12083701 06/3180 5 4 3 2 1 0

Lecture Notes in Computer Science 4592

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Preface

The 12th International Conference on Applications of Natural Language to Information Systems (NLDB 2007) took place during June 27–29 in Paris (France). Since the first edition in 1995, the NLDB conference has been aiming at bringing together researchers, people working in industry and potential users interested in various applications of natural language in the database and information system areas.

Natural language and databases are core components in the development of information systems. NLP techniques may substantially enhance most phases of the information system lifecycle, starting with requirement analysis, specification and validation, and going up to conflict resolution, result processing and presentation. Furthermore, natural language-based query languages and user interfaces facilitate the access to information for all and allow for new paradigms in the usage of computerized services. Hot topics such as information retrieval and Semantic Web-based applications imply a complete fusion of databases and NLP techniques.

Among an increasing number of submitted papers (110), the Program Committee selected 31 papers as full papers, thus coming up with an acceptance rate of 28%. These proceedings also include 12 short papers that were presented at the conference and two invited talks, one given by Andrew Basden and Heinz Klein and the other given by Max Silberztein.

This conference was possible thanks to the support of three organizing institutions: The University of Versailles Saint-Quentin (Versailles, France), the Conservatoire National des Arts et Métiers (Paris, France) and the University of Salford (Salford, UK). We thank them, and in particular Profs. Akoka (CNAM), Bouzeghoub (UVSQ) and Comyn-Wattiau (CNAM) for their support.

We also wish to thank the entire organizing team including secretaries, researchers and students who put their competence, enthusiasm and kindness into making this conference a real success, and especially Xiaohui Xue, who managed the Web site.

June 2007

Zoubida Kedad
Nadira Lammari
Elisabeth Métais
Farid Meziane
Yacine Rezgui

Conference Organization

Conference Co-chairs

Elisabeth Metais
Jacky Akoka
Mokrane Bouzeghoub
Yacine Rezgui

Program Co-chairs

Zoubida Kedad
Isabelle Comyn-Wattiau
Farid Meziane

Organization Chair

Nadira Lammari

Program Committee

Witold Abramowicz, The Poznañ University of Economics, Poland
Frederic Andres, University of Advanced Studies, Japan
Kenji Araki, Hokkaido University, Japan
Akhilesh Bajaj, University of Tulsa, USA
Maria Bergoltz, Stockholm University, Sweden
Marc El-Beze, CNRS Laboratoire Informatique d'Avignon, France
Béatrice Bouchou, Université François-Rabelais de Tours, France
Mokrane Bouzeghoub, Université de Versailles, France
Andrew Burton-Jones, University of British Columbia, Canada
Hiram Calvo, National Polytechnic Institute, Mexico
Roger Chiang, University of Cincinnati, USA
Gary A Coen, Boeing, USA
Isabelle Comyn-Wattiau, CNAM, France
Cedric Du Mouza, CNAM, France
Antje Düsterhöft, University of Wismar, Germany
Günther Fliedl, Universität Klagenfurt, Austria
Christian Fluhr, CEA, France
Alexander Gelbukh, Instituto Politecnico Nacional, Mexico
Jon Atle Gulla, Norwegian University of Science and Technology, Norway
Udo Hahn, Friedrich-Schiller-Universität Jena, Germany
Karin Harbusch, Universität Koblenz-Landau, Germany

Harmain Harmain, United Arab Emirates University, UAE
Helmut Horacek, Universität des Saarlandes, Germany
Cecil Chua Eng Huang, Nanyang Technological University, Singapore
Paul Johannesson, Stockholm University, Sweden
Epaminondas Kapetanios, University of Westminster, UK
Asanee Kawtrakul, Kasetsart University, Thailand
Zoubida Kedad, Université de Versailles, France
Christian Kop, University of Klagenfurt, Austria
Leila Kosseim, Concordia University, Canada
Nadira Lammari, CNAM, France
Winfried Lenders, Universität Bonn, Germany
Jana Lewerenz, sd&m Düsseldorf, Germany
Deryle Lonsdale, Brigham Young University, USA
Stéphane Lopes, Université de Versailles, France
Robert Luk, Hong Kong Polytechnic University, Hong Kong
Bernardo Magnini, IRST, Italy
Heinrich C. Mayr, University of Klagenfurt, Austria
Paul McFetridge, Simon Fraser University, Canada
Elisabeth Metais, CNAM , France
Farid Meziane, Salford University, UK
Luisa Mich, University of Trento, Italy
Ruslan Mitkov, University of Wolverhampton, UK
Diego Mollá Aliod, Macquarie University, Australia
Andrés Montoyo, Universidad de Alicante, Spain
Ana Maria Moreno, Universidad Politecnica de Madrid, Spain
Rafael Muñoz, Universidad de Alicante, Spain
Samia Nefti-Meziani, Salford University, UK
Günter Neumann, DFKI, Germany
Jian-Yun Nie, Université de Montréal, Canada
Manual Palomar, Universidad de Alicante, Spain
Pit Pichappan, Annamalai University, India
Odile Piton, Université Paris I Panthéon-Sorbonne, France
Violaine Prince, Université Montpellier 2/LIRMM-CNRS, France
Sandeep Purao, Pennsylvania State University, USA
Yacine Rezgui, University of Salford, UK
Reind van de Riet, Vrije Universiteit Amsterdam, The Netherlands
Hae-Chang Rim, Korea University, Korea
Samira si-Said, CNAM, France
Grigori Sidorov, Instituto Politecnico Nacional, Mexico
Max Silberztein, Université de Franche-Comté, France
Veda Storey, Georgia State University, USA
Vijayan Sugumaran, Oakland University Rochester, USA
Lua Kim Teng, National University of Singapore, Singapore
Bernhard Thalheim, Kiel University, Germany
Krishnaprasad Thirunarayan, Wright State University, USA
Juan Carlos Trujillo, Universidad de Alicante, Spain
Luis Alfonso Ureña, Universidad de Jaén, Spain

Panos Vassiliadis, University of Ioannina, Greece
Jürgen Vöhringer, University of Klagenfurt, Austria
Roland Wagner, University of Linz, Austria
Hans Weigand, Tilburg University, The Netherlands
Werner Winiwarter, University of Vienna, Austria
Christian Winkler, Universität Klagenfurt, Austria
Stanislaw Wrycza, University of Gdansk, Poland

Additional Reviewers

Jing Bai, Norman Biehl, Terje Brasethevik, Gaël De Chalendar, Hiroshi Echizen-Ya, Óscar Ferrandez, Miguel A.Garcia, Gregory Grefenstette, Trivikram Immaneni, Jon Espen Ingvaldsen, Zornitsa Kozareva, Teresa Martin, Fernando Martinez-Santiago, Arturo Montejo-Raez, Borja Navarro, Octavian Popescu, Rafal Rzepka, Agata Savary, Hideyuki Shibuki, Jonas Sjobergh, Darijus Strasunskas, David Tomas, Stein L.Tomassen

Lecture Notes in Computer Science

For information about Vols. 1–4476

please contact your bookseller or Springer

Vol. 4600: H. Comon-Lundh, C. Kirchner, H. Kirchner (Eds.), Rewriting, Computation and Proof. XVI, 273 pages. 2007.

Vol. 4592: Z. Kedad, N. Lammari, E. Métais, F. Meziane, Y. Rezgui (Eds.), Natural Language Processing and Information Systems. XIV, 442 pages. 2007.

Vol. 4591: J. Davies, J. Gibbons (Eds.), Integrated Formal Methods. IX, 660 pages. 2007.

Vol. 4590: W. Damm, H. Hermanns (Eds.), Computer Aided Verification. XV, 562 pages. 2007.

Vol. 4588: T. Harju, J. Karhumäki, A. Lepistö (Eds.), Developments in Language Theory. XI, 423 pages. 2007.

Vol. 4584: N. Karssemeijer, B. Lelieveldt (Eds.), Information Processing in Medical Imaging. XIII, 775 pages. 2007.

Vol. 4583: S.R. Della Rocca (Ed.), Typed Lambda Calculi and Applications. XI, 395 pages. 2007.

Vol. 4581: A. Petrenko, M. Veanes, J. Tretmans, W. Grieskamp (Eds.), Testing of Software and Communicating Systems. XII, 379 pages. 2007.

Vol. 4574: J. Derrick, J. Vain (Eds.), Formal Techniques for Networked and Distributed Systems – FORTE 2007. XI, 375 pages. 2007.

Vol. 4573: M. Kauers, M. Kerber, R. Miner, W. Windsteiger (Eds.), Towards Mechanized Mathematical Assistants. XIII, 407 pages. 2007. (Sublibrary LNAI).

Vol. 4572: F. Stajano, C. Meadows, S. Capkun, T. Moore (Eds.), Security and Privacy in Ad-hoc and Sensor Networks. X, 247 pages. 2007.

Vol. 4569: A. Butz, B. Fisher, A. Krüger, P. Olivier, S. Owada (Eds.), Smart Graphics. IX, 237 pages. 2007.

Vol. 4549: J. Aspnes, C. Scheideler, A. Arora, S. Madden (Eds.), Distributed Computing in Sensor Systems. XIII, 417 pages. 2007.

Vol. 4548: N. Olivetti (Ed.), Automated Reasoning with Analytic Tableaux and Related Methods. X, 245 pages. 2007. (Sublibrary LNAI).

Vol. 4547: C. Carlet, B. Sunar (Eds.), Arithmetic of Finite Fields. XI, 355 pages. 2007.

Vol. 4546: J. Kleijn, A. Yakovlev (Eds.), Petri Nets and Other Models of Concurrency – ICATPN 2007. XI, 515 pages. 2007.

Vol. 4543: A.K. Bandara, M. Burgess (Eds.), Inter-Domain Management. XII, 237 pages. 2007.

Vol. 4542: P. Sawyer, B. Paech, P. Heymans (Eds.), Requirements Engineering: Foundation for Software Quality. IX, 384 pages. 2007.

Vol. 4541: T. Okadome, T. Yamazaki, M. Makhtari (Eds.), Pervasive Computing for Quality of Life Enhancement. IX, 248 pages. 2007.

Vol. 4539: N.H. Bshouty, C. Gentile (Eds.), Learning Theory. XII, 634 pages. 2007. (Sublibrary LNAI).

Vol. 4538: F. Escolano, M. Vento (Eds.), Graph-Based Representations in Pattern Recognition. XII, 416 pages. 2007.

Vol. 4537: K.C.-C. Chang, W. Wang, L. Chen, C.A. Ellis, C.-H. Hsu, A.C. Tsoi, H. Wang (Eds.), Advances in Web and Network Technologies, and Information Management. XXIII, 707 pages. 2007.

Vol. 4536: G. Concas, E. Damiani, M. Scotto, G. Succi (Eds.), Agile Processes in Software Engineering and Extreme Programming. XV, 276 pages. 2007.

Vol. 4534: I. Tomkos, F. Neri, J. Solé Pareta, X. Masip Bruin, S. Sánchez Lopez (Eds.), Optical Network Design and Modeling. XI, 460 pages. 2007.

Vol. 4531: J. Indulska, K. Raymond (Eds.), Distributed Applications and Interoperable Systems. XI, 337 pages. 2007.

Vol. 4530: D.H. Akehurst, R. Vogel, R.F. Paige (Eds.), Model Driven Architecture- Foundations and Applications. X, 219 pages. 2007.

Vol. 4529: P. Melin, O. Castillo, L.T. Aguilar, J. Kacprzyk, W. Pedrycz (Eds.), Foundations of Fuzzy Logic and Soft Computing. XIX, 830 pages. 2007. (Sublibrary LNAI).

Vol. 4528: J. Mira, J.R. Álvarez (Eds.), Nature Inspired Problem-Solving Methods in Knowledge Engineering, Part II. XXII, 650 pages. 2007.

Vol. 4527: J. Mira, J.R. Álvarez (Eds.), Bio-inspired Modeling of Cognitive Tasks, Part I. XXII, 630 pages. 2007.

Vol. 4526: M. Malek, M. Reitenspieß, A. van Moorsel (Eds.), Service Availability. X, 155 pages. 2007.

Vol. 4525: C. Demetrescu (Ed.), Experimental Algorithms. XIII, 448 pages. 2007.

Vol. 4524: M. Marchiori, J.Z. Pan, C.d.S. Marie (Eds.), Web Reasoning and Rule Systems. XI, 382 pages. 2007.

Vol. 4523: Y.-H. Lee, H.-N. Kim, J. Kim, Y. Park, L.T. Yang, S.W. Kim (Eds.), Embedded Software and Systems. XIX, 829 pages. 2007.

Vol. 4522: B.K. Ersbøll, K.S. Pedersen (Eds.), Image Analysis. XVIII, 989 pages. 2007.

Vol. 4521: J. Katz, M. Yung (Eds.), Applied Cryptography and Network Security. XIII, 498 pages. 2007.

Vol. 4519: E. Franconi, M. Kifer, W. May (Eds.), The Semantic Web: Research and Applications. XVIII, 830 pages. 2007.

Vol. 4517: F. Boavida, E. Monteiro, S. Mascolo, Y. Koucheryavy (Eds.), Wired/Wireless Internet Communications. XIV, 382 pages. 2007.

Vol. 4516: L. Mason, T. Drwiega, J. Yan (Eds.), Managing Traffic Performance in Converged Networks. XXIII, 1191 pages. 2007.

Vol. 4515: M. Naor (Ed.), Advances in Cryptology - EUROCRYPT 2007. XIII, 591 pages. 2007.

Vol. 4514: S.N. Artemov, A. Nerode (Eds.), Logical Foundations of Computer Science. XI, 513 pages. 2007.

Vol. 4513: M. Fischetti, D.P. Williamson (Eds.), Integer Programming and Combinatorial Optimization. IX, 500 pages. 2007.

Vol. 4511: C. Conati, K. McCoy, G. Paliouras (Eds.), User Modeling 2007. XVI, 487 pages. 2007. (Sublibrary LNAI).

Vol. 4510: P. Van Hentenryck, L. Wolsey (Eds.), Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems. X, 391 pages. 2007.

Vol. 4509: Z. Kobti, D. Wu (Eds.), Advances in Artificial Intelligence. XII, 552 pages. 2007. (Sublibrary LNAI).

Vol. 4508: M.-Y. Kao, X.-Y. Li (Eds.), Algorithmic Aspects in Information and Management. VIII, 428 pages. 2007.

Vol. 4507: F. Sandoval, A. Prieto, J. Cabestany, M. Graña (Eds.), Computational and Ambient Intelligence. XXVI, 1167 pages. 2007.

Vol. 4506: D. Zeng, I. Gotham, K. Komatsu, C. Lynch, M. Thurmond, D. Madigan, B. Lober, J. Kvach, H. Chen (Eds.), Intelligence and Security Informatics: Biosurveillance. XI, 234 pages. 2007.

Vol. 4505: G. Dong, X. Lin, W. Wang, Y. Yang, J.X. Yu (Eds.), Advances in Data and Web Management. XXII, 896 pages. 2007.

Vol. 4504: J. Huang, R. Kowalczyk, Z. Maamar, D. Martin, I. Müller, S. Stoutenburg, K.P. Sycara (Eds.), Service-Oriented Computing: Agents, Semantics, and Engineering. X, 175 pages. 2007.

Vol. 4501: J. Marques-Silva, K.A. Sakallah (Eds.), Theory and Applications of Satisfiability Testing – SAT 2007. XI, 384 pages. 2007.

Vol. 4500: N. Streitz, A. Kameas, I. Mavrommati (Eds.), The Disappearing Computer. XVIII, 304 pages. 2007.

Vol. 4499: Y.Q. Shi (Ed.), Transactions on Data Hiding and Multimedia Security II. IX, 117 pages. 2007.

Vol. 4498: N. Abdennahder, F. Kordon (Eds.), Reliable Software Technologies – Ada Europe 2007. XII, 247 pages. 2007.

Vol. 4497: S.B. Cooper, B. Löwe, A. Sorbi (Eds.), Computation and Logic in the Real World. XVIII, 826 pages. 2007.

Vol. 4496: N.T. Nguyen, A. Grzech, R.J. Howlett, L.C. Jain (Eds.), Agent and Multi-Agent Systems: Technologies and Applications. XXI, 1046 pages. 2007. (Sublibrary LNAI).

Vol. 4495: J. Krogstie, A. Opdahl, G. Sindre (Eds.), Advanced Information Systems Engineering. XVI, 606 pages. 2007.

Vol. 4494: H. Jin, O.F. Rana, Y. Pan, V.K. Prasanna (Eds.), Algorithms and Architectures for Parallel Processing. XIV, 508 pages. 2007.

Vol. 4493: D. Liu, S. Fei, Z. Hou, H. Zhang, C. Sun (Eds.), Advances in Neural Networks – ISNN 2007, Part III. XXVI, 1215 pages. 2007.

Vol. 4492: D. Liu, S. Fei, Z. Hou, H. Zhang, C. Sun (Eds.), Advances in Neural Networks – ISNN 2007, Part II. XXVII, 1321 pages. 2007.

Vol. 4491: D. Liu, S. Fei, Z.-G. Hou, H. Zhang, C. Sun (Eds.), Advances in Neural Networks – ISNN 2007, Part I. LIV, 1365 pages. 2007.

Vol. 4490: Y. Shi, G.D. van Albada, J. Dongarra, P.M.A. Sloot (Eds.), Computational Science – ICCS 2007, Part IV. XXXVII, 1211 pages. 2007.

Vol. 4489: Y. Shi, G.D. van Albada, J. Dongarra, P.M.A. Sloot (Eds.), Computational Science – ICCS 2007, Part III. XXXVII, 1257 pages. 2007.

Vol. 4488: Y. Shi, G.D. van Albada, J. Dongarra, P.M.A. Sloot (Eds.), Computational Science – ICCS 2007, Part II. XXXV, 1251 pages. 2007.

Vol. 4487: Y. Shi, G.D. van Albada, J. Dongarra, P.M.A. Sloot (Eds.), Computational Science – ICCS 2007, Part I. LXXXI, 1275 pages. 2007.

Vol. 4486: M. Bernardo, J. Hillston (Eds.), Formal Methods for Performance Evaluation. VII, 469 pages. 2007.

Vol. 4485: F. Sgallari, A. Murli, N. Paragios (Eds.), Scale Space and Variational Methods in Computer Vision. XV, 931 pages. 2007.

Vol. 4484: J.-Y. Cai, S.B. Cooper, H. Zhu (Eds.), Theory and Applications of Models of Computation. XIII, 772 pages. 2007.

Vol. 4483: C. Baral, G. Brewka, J. Schlipf (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 327 pages. 2007. (Sublibrary LNAI).

Vol. 4482: A. An, J. Stefanowski, S. Ramanna, C.J. Butz, W. Pedrycz, G. Wang (Eds.), Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. XIV, 585 pages. 2007. (Sublibrary LNAI).

Vol. 4481: J. Yao, P. Lingras, W.-Z. Wu, M. Szczuka, N.J. Cercone, D. Ślęzak (Eds.), Rough Sets and Knowledge Technology. XIV, 576 pages. 2007. (Sublibrary LNAI).

Vol. 4480: A. LaMarca, M. Langheinrich, K.N. Truong (Eds.), Pervasive Computing. XIII, 369 pages. 2007.

Vol. 4479: I.F. Akyildiz, R. Sivakumar, E. Ekici, J.C.d. Oliveira, J. McNair (Eds.), NETWORKING 2007. Ad Hoc and Sensor Networks, Wireless Networks, Next Generation Internet. XXVII, 1252 pages. 2007.

Vol. 4478: J. Martí, J.M. Benedí, A.M. Mendonça, J. Serrat (Eds.), Pattern Recognition and Image Analysis, Part II. XXVII, 657 pages. 2007.

Vol. 4477: J. Martí, J.M. Benedí, A.M. Mendonça, J. Serrat (Eds.), Pattern Recognition and Image Analysis, Part I. XXVII, 625 pages. 2007.

Table of Contents

Ontology Engineering

Natural Language for Information System Design

Information Retrieval Systems

Natural Language Processing Techniques

Short Papers

An Alternative Approach to Tagging

Max Silberztein

LASELDI, Université de Franche-Comté
30 rue Mégevand, 25000 Besançon, France
max.silberztein@univ-fcomte.frUT

Abstract. NooJ is a linguistic development environment that allows users to construct large formalised dictionaries and grammars and use these resources to build robust NLP applications. NooJ's approach to the formalisation of natural languages is bottom-up: linguists start by formalising basic phenomena such as spelling and morphology, and then formalise higher and higher linguistic levels, moving up towards the sentence level. NooJ provides parsers that operate in cascade at each individual level of the formalisation: tokenizers, morphological analysers, simple and compound terms indexers, disambiguation tools, syntactic parsers, named entities annotators and semantic analysers. This architecture requires NooJ's parsers to communicate via a Text Annotation Structure that stores both correct results and erroneous hypotheses (to be deleted later).

Keywords: NooJ. Linguistic Development Environment. Robust NLP applications.

1 Introduction

NooJ is a linguistic development environment that allows users to construct large linguistic resources in the form of electronic dictionaries and grammars and to apply these resources to large texts and build various robust NLP applications.[1]

NooJ's approach to the formalisation of natural languages is bottom-up: linguists start by formalising basic phenomena, such as spelling, morphology and lexicon, and then use these basic levels of description to formalise higher and higher linguistic levels, moving up towards the syntactic and semantic levels. This bottom-up approach is complemented by an accumulative methodology that allows a community of users to share and re-use individual resources, as well as a number of tools (e.g. concordances, contract enforcers and debuggers) that help users maintain the integrity of large resources.

Parallel to these various levels of formalisation, NooJ provides a number of parsers that operate in cascade at each individual level of the formalisation: at the character level (tokenizer and sentence recogniser), morphology (inflectional and derivational analyser), lexical (simple and compound words recognisers), local syntax (disambiguation), structural syntax (frozen and semi-frozen expressions recogniser, syntactic parser) and transformational syntax (semantic analyser).

[1] NooJ is freeware. See: http://www.nooj4nlp.netUT to download the software and its documentation.

Z. Kedad et al.(Eds.): NLDB 2007, LNCS 4592, pp. 1–11, 2007.

2 A Third Type of NLP Tool

Most available NLP tools follow one of two different and incompatible approaches.

On the one hand, some linguistic parsers aim at formalising natural languages, usually at the syntactic and semantic levels. Following Chomsky's discussion of the inadequacies of finite-state machines for NLP [1], researchers have invented and refined several computational devices and their corresponding formalisms capable of representing complex, non finite-state syntactic phenomena, such as unification-based parsers that deal with various types of agreement constraints.

Unfortunately most of these parsers, while powerful enough to compute a variety of complex syntactic analyses, are not adapted to the processing of very simple but cost-intensive phenomena, such as locating multi-word expressions in texts by accessing a dictionary of over 200,000 entries, performing morphological analysis of Hungarian texts, etc. Moreover, they are not capable of parsing large corpora in real-time, and therefore cannot be used as online corpus processing tools, nor can they be used as linguistic engines for "basic" applications such as search engines.

On the other hand, some NLP tools aim at facilitating the implementation of NLP applications such as search engines, automatic construction of abstracts, corpus processors, information extraction, etc. These tools often include very efficient parsers based on finite-state technology, and can indeed be used to parse large quantities of texts. Unfortunately, these tools include at one point or another several algorithms that make them unsuitable to the formalisation of natural languages, such as a statistical tagger that aims at producing "reasonably good" results – which is to say a number of incorrect ones – as well as heuristics to get rid of ambiguities – even when sentences are genuinely ambiguous – etc.

NooJ shares with the above-mentioned linguistic tools the goal of providing linguists with a way to formalise natural languages precisely, and at the same time includes several efficient finite-state tools to parse large texts and process large linguistic resources. This can be done because in NooJ's bottom-up architecture, each level of analysis is processed by a different, specialised (and therefore efficient) computational device. In other words, instead of using one single powerful (and inefficient) computational device to process all kinds of linguistic phenomena, we assume that natural languages are sets of very different phenomena, each of them requiring a specialized mechanism and associated parser; in particular, simple devices such as finite-state machines, which constitute very natural tools to represent a large number of linguistic phenomena, should not be thrown away because they are not adequate to represent other, sometimes even exotic, phenomena.

3 Atomic Linguistic Units

In NooJ, the term **Atomic Linguistic Units** (ALUs) refers to the smallest elements of a given language that are associated with linguistic information. By definition, these ALUs constitute the vocabulary of the language. They can and must be systematically

described in extension, because some of, or all their properties cannot be computed from their components.

NooJ's first level of text analysis is at the character level. Characters are classified as **letters** and **delimiters**. **Tokens** are sequences of letters between delimiters. Based on these three definitions, NooJ distinguishes four types of ALUs:

- **Simple Word**: any ALU spelled as a token, e.g. *table*
- **Affix**: any ALU spelled as a subsequence of letters in a token, e.g. *re-, -able*
- **Multi-Word Unit** (MWU): any ALU spelled as a sequence of letters and delimiters, e.g. *as a matter of fact* (the space character is not a letter, hence it is a delimiter)
- **Frozen Expression**: any MWU that accepts insertions, e.g. *take ... into account*

This classification is adequate for any written language, although some languages (e.g. English) require the description of very few affixes whereas others (e.g. Hungarian) require a large morphological system. The vocabulary of Romance languages probably contains five times more MWUs than simple words, whereas Germanic languages have very few MWUs because these units are not spelled with spaces, and hence are processed by NooJ as simple words.[2]

Obviously, it is important for any NLP application to be able to process any type of ALUs, including MWUs. Questions about MWUs – their definition, how to automatically recognise them in texts, how to process them, etc. – have initiated much interest in computational linguistics. Unfortunately, there seems to be confusion about what people mean by MWUs (or compounds, or complex terms, collocations, etc.). For instance, [2] studies "transparent" noun compounds and discusses methods for analysing them. In the NooJ framework, this is contradictory: either these objects are truly MWUs (i.e. **Atomic** Linguistic Units) that must be listed and explicitly described in dictionaries (and therefore don't need to be analysed), or they are transparent and can be analysed automatically (and therefore do not need to be listed explicitly in a dictionary). Note finally that more precise analyses (such as [3]) show that so-called "transparent" MWUs have in fact some degrees of "opacity" that would require robust NLP application to list and describe them explicitly in a dictionary.

Statisticians often equate compounds with collocations, i.e. sequences of tokens that occur together frequently. But trying to characterise MWUs by computing the co-occurrence of their components negates their atomicity.[3] Moreover, MWUs (just like simple words) can be either frequent or rare in texts, and MWUs with a low frequency (say, less than 3 occurrences in a large corpus), are typically left out by statistical methods.

---

[2] On the other hand, Germanic analysable tokens, such as "*Schiffahrtsgesellschaft*" must be processed as sequences of affixes: "*Schiff fahrt s gesellschaft*".

[3] In the same manner that one should not try to prove the fact that the token "apartment" is an English word from the fact that "apart" and "ment" often occurs together. In fact, tokenizers used by most taggers and statistical parsers naively use the blank character to characterise linguistic units. In NooJ, MWUs are ALUs just like simple words; the fact that they include blanks or other delimiters is not relevant to their status, and does not even complicate their automatic recognition in texts.