A lexicon for Vietnamese language processing

Nguyen T.M.H., Romary L., Rossignol M., Vu X.L.
Faculty of Mathematics, Mechanics and Informatics, Hanoi University of Science, 334 Nguyen Trai, Hanoi, 10000, Viet Nam; LORIA, Nancy, France; International Research Center MICA, Hanoi, Viet Nam; Vietnam Lexicography Center, Hanoi, Viet Nam

Abstract: Only very recently have Vietnamese researchers begun to be involved in the domain of Natural Language Processing (NLP). As there does not exist any published work in formal linguistics nor any recognizable standard for Vietnamese word definition and word categories, the fundamental tasks for automatic Vietnamese language processing, such as part-of-speech tagging, parsing, etc., are very difficult tasks for computer scientists. The fact that all necessary linguistic resources have to be built from scratch by each research team is a real obstacle to the development of Vietnamese language processing. The aim of our projects is thus to build a common linguistic database that is freely and easily exploitable for the automatic processing of Vietnamese. In this paper, we present our work on creating a Vietnamese lexicon for NLP applications. We emphasize the standardization aspect of the lexicon representation. We especially propose an extensible set of Vietnamese syntactic descriptions that can be used for tagset definition and morphosyntactic analysis. These descriptors are established in such a way as to be a reference set proposal for Vietnamese in the context of ISO subcommittee TC 37/SC 4 (Language Resource Management).
Author Keywords: Lexicon; Linguistic resources; Part-of-speech; Standardization; Syntactic description; Vietnamese

Year: 2006
Source title: Language Resources and Evaluation
Volume: 40
Issue: 4-Mar
Page : 291-309
Link: Scorpus Link
Correspondence Address: Nguyễn, T.M.H.; Faculty of Mathematics, Mechanics and Informatics, Hanoi University of Science, 334 Nguyen Trai, Hanoi, 10000, Viet Nam; email: huyenntm@vnu.edu.vn
ISSN: 1574020X
DOI: 10.1007/s10579-007-9034-8
Language of Original Document: English
Abbreviated Source Title: Language Resources and Evaluation
Document Type: Article
Source: Scopus
Authors with affiliations:

- Nguyễn, T.M.H., Faculty of Mathematics, Mechanics and Informatics, Hanoi University of Science, 334 Nguyen Trai, Hanoi, 10000, Viet Nam

- Romary, L., LORIA, Nancy, France
- Rossignol, M., International Research Center MICA, Hanoi, Viet Nam
- Vũ, X.L., Vietnam Lexicography Center, Hanoi, Viet Nam

References:

- Cao, X.H., (2000) Tiêng Viát-mây Vân Crossed D Signê Âm, Ngá Nghĩa (Vietnamese-Some Questions on Phonetics, Syntax and Semantics), , NXB Giào dác Hà Nái, Viäët Nam
- Dien, D., Hoi, P.P., Hung, N.Q., Some lexical issues in electronic Vietnamese dictionary (2003) PAPILLON-2003 Workshop on Multilingual Lexical Databases, , Hokaido University, Japan
- Dien, D., Kiem, H., POS-tagger for English-Vietnamese bilingual corpus (2003) Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and beyond, , Canada: Edmonton
- Dien, D., Kiem, H., State of the art of machine translation in Vietnam (2005) AAMT Journal, Special Issue on MT Summit X
- Dien, D., Kiem, H., Toan, N.V., Vietnamese word segmentation (2001) Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001), , Tokyo, Japan
- Diáp, Q.B., Văn Thung, H., (1999) Ngá Pháp Tiêng Viát (Vietnamese Grammar), 1. , NXB Giào dác Hà Nái, Viäët Nam
- Erjavec, T., Ide, N., Tufis, D., Development and assessment of common lexical specifications for six central and eastern European languages (1998) Proceedings of the First International Conference on Language Resources and Evaluation, , Granada, Spain
- Hoàng, P., (2002) Tà Crossed D Signien Tiêng Viát (Vietnamese Dictionary), , NXB Crossed D signà Naong Viäët Nam
- Háu Sign., C.D., Dõi, T.T., Lan Sign. T, C.D., (1998) Csaytiêng Viát (Basis of Vietnamese), , NXB Giào dác Hà Nái, Viäët Nam
- Ide, N., Romary, L., Standards for language resources (2001) Proceedings of the IRCS Workshop on Linguistic Databases, , Philadelphia, US
- Ide, N., Romary, L., Abeillè, A., Encoding syntactic annotation (2003) Building and Using Parsed Corpora, , Kluwer Academic Publishers Dordrecht, Netherlands
- Ide, N., Véronis, J., MULTEXT: Multilingual text tools and corpora (1994) Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), , Kyoto, Japan
- Ide, N., Véronis, J., Ide, N., Véronis, J., Encoding dictionaries (1995) Text Encoding Initiative: Background and Context, , Kluwer Academic Publishers Dordrecht, Netherlands
- (2006) Language Resource Management-Lexical Markup Framework (LMF), , ISO 24613, Rev.13 ISO, Geneva, Switzerland
- Li, C.N., Thompson, S.A., Li, C.N., Subject and topic: A new typology of language (1976) Subject and Topic, pp. 457-489. , Academic Press London/New York
- Nguyen, T.M.H., Romary, L., Vu, X.L., Une étude de cas pour l'étiquetage morpho-syntaxique de textes Vietnamiens (2003) Actes de la Conférence Francophone Internationale sur le Traitement Automatique des Langues Naturelles (TALN 03), , Batz-sur-mer, France
- Nguyễn, T.M.H., Outils et ressources linguistiques pour l'alignement de textes multilingues Français-Vietnamiens (2006) Thèse de Doctorat en Informatique, , Université Henri Poincaré, Nancy I, Nancy, France
- Nguyễn, T.C., (1998) Ngá Pháp Tiêng Viát (Vietnamese Grammar), , NXB Crossed D signai hac Quôc gia Hà Nái, Viäët Nam
- Romary, L., Salmon-Alt, S., Francopoulo, G., Standards going concrete: From LMF to Morphalou (2004) The 20th International Conference on Computational Linguistics (COLING), , Workshop Enhancing and using electronic dictionaries Geneva, Switzerland
- (1983) Ngá Pháp Tiêng Viát (Vietnamese Grammar), , Uában Khoa hac Xã hái Viát Nam Hà Nái, Viát Nam: NXB Khoa hac Xã

hái