ENHANCING CITIZEN-GOVERNMENT COMMUNICATION THROUGH NATURAL LANGUAGE QUERYING

Lawrence Muchemi, Getao Katherine

School of Computing and Informatics
University of Nairobi
{lmuchemi, kgetao}@uonbi.ac.ke

ABSRACT. The fast growth of computerized data processing and communications, and the desire by most African governments to introduce e-governance has necessitated efficient dissemination of public sector data. A key challenge in the implementation of these systems will be data retrieval from relational databases that store public data. Most members of the general public in East Africa are not conversant with electronic database query techniques and therefore may not take full advantage of computerized government information systems unless user-friendly interfaces are provided to assist them. In this paper we propose Swahili natural language querying (NLQ) as an effective means of solving this problem. However there are some key unresolved issues within Swahili NLQ that must be addressed in order to effect this solution.

We summarize the current theoretical and methodological approaches being applied to Swahili natural language querying (NLQ) and identify the key outstanding research issues. These issues include lack of computerized phrase identification mechanisms, difficulties in computerized information extraction using traditional syntactic processing methods and the lack of algorithms that translate Swahili NL queries into Structured Query Language (SQL.) We propose a solution that utilizes a frame-based semantic representation of Swahili phrases.

Keywords: natural language querying (NLQ), meaning representations, frames

1 Introduction

Many governments in developing countries struggle with the problem of inefficient communication with their citizenry. The public has minimal access to information buried within huge data repositories. For example, a key objective in the Kenya's e-government strategy and standardization framework [5] is to enable citizens to access Government services and information as efficiently and effectively as possible through the use of the Internet and other channels of communication.

A public communication strategy within an E-government framework should ideally allow citizens to directly access public information stored in government databases. Current database querying techniques rely on command line interfaces or data entry form-based interfaces that structure appropriate SQL queries. The former require technically adept users while the latter exhibit the inherent limitations of template-based input formats. The ideal query mechanism would enable the users to seek information from the database in the same way that they would question a human officer: using natural language. This paper therefore proposes the use of natural language (NL) text-based database querying system using Swahili, the most commonly used language in Eastern Africa, to access public data and information.

Natural language processing has experienced a new impetus in recent years with many researchers devoting attention to natural language processing for Swahili. This research

has created a substantial body of knowledge and expertise in this domain including [7, 13, 14, 18] among others.

2.0 OUTSTANDING RESEARCH ISSUES

2.1 Important research issues for Swahili natural language database query systems

We aim to solve the problem of user-friendly access to public data by developing a Swahili Natural Language Database Query system (SNLDQS.) To do this effectively we must apply existing knowledge in Swahili natural language processing within the solution. During our study of the existing domain knowledge we have identified some critical gaps that must be addressed in order to succeed in developing the SNLDQS. These include

- § Lack of automatic phrase recognition mechanisms for Swahili texts,
- § Difficulties in automatic phrase extraction using traditional syntactic processing methods, and
- § The lack of algorithms that translate Swahili natural language phrases into SQL queries.

An extensive literature search led to few references in Swahili natural language database access. However there is robust research in this area that addresses other languages. This body of knowledge coupled with a growing population of researchers in Swahili natural language processing has encouraged the authors that this technology can be extended to encompass the Swahili language.

2.1.1 Automatic phrase recognition mechanisms

In almost all natural languages, words and phrases are used to carry meanings. When we have words grouped together to represent concepts in specified domains, we refer to this grouping as a phrase. The rules, also known as productions, guide these formations are language specific. The process of identifying these productions requires the identification of categories of words in a language. Such categories include noun classes (ngeli for Swahili), verb classes (vitenzi), connectors (viunganishi) etc. [20]. Once these word categories are identified the rules that guide the production of phrases can be manually studied by linguists. Temu [19] used this method to obtain six formation patterns for Swahili phrases

The corpus-based approach is an alternative methodology for the discovery of word classes and phrase rules. This methodology induces information from huge collections of natural language texts, known as corpora. The induction process may be manual or automated.. Ohly [15] applied the corpus based approach to textile domain and obtained six patterns including:

i.) **VN N** (Norminalized verb phrase) {for VN the verb is being used as a Noun }
Example, Ku-kaza Uzi

ii.) **DV N** Deverbative head with a noun complement
Example Ki-onja Mchuzi {verbs converted to nouns}

iii.) **N N** (construction of two nouns in juxta position
Example eg Haidrogeni peroksaidi)

iv.) **N Adj** (combination of noun and adjective)
Example Ndovu Mweupe, almasi nyeusi

v.) **N –a N (**Construction with a connector –a
Example mizani y-a bondia; Chakula ch-a motto others include 'z-a', 'w-a', l-a etc

vi.) **N -a VN** (Construction with connector –a followed by verb noun qualifier)
Example Gari la Kusafiria;
(Sewangi 2002: 27).

Such phrase structure grammars have previously been in existence for Swahili in the form of context-free grammars (CFGs) [20]. However no algorithms or other computer-based mechanisms for automatic phrase recognition based on these or other patterns have been developed. This is a key research gap that is addressed in this paper. Phrase recognition is a precursor to phrase extraction, which can be achieved by phrase pattern search followed by comparison with a lexical database. This method is likely to be more efficient than key word(s) search reported in Kitani and Eriguchi [12].

2.1.2 Automatic phrase extraction

Most sources reviewed revealed that extraction of terms from natural language texts relies on deep syntactic processing or deep semantic processing. Each method has advantages and disadvantages.

Syntactic processing has the potential to reveal details from a sentence but cannot handle sentences that are grammatically incorrect yet can be understood by people. For example:

Mimi iko taka maji. {me want water}

Such sentences cannot be successfully syntactically parsed and hence a phrase extraction algorithm that relies solely on syntactic processing will be unable to process this sentence. This phenomenon of system failure due to changes in language is known as 'brittleness'.

Deep semantic processing on the other hand is guided by specialized domain knowledge that enables the system to overcome minor syntactic variations. However, this has the disadvantage of creating inflexible systems that are difficult to port to other knowledge domains. A new semantic grammar has to be written whenever the system is configured for a new knowledge domain. Thus this approach of information extraction has been slow to yield the desired results. However, the problem of improving the portability of semantic phrase extraction systems remains a popular research topic among NLQ researchers across the globe.

In this paper we report on a hybrid approach which combines shallow syntactic processing (to identify phrases) with shallow semantic processing through mapping of meaning representations (MR).

2.1.3 Algorithms that translate Swahili phrases into SQL queries

There is currently limited research in the area of Swahili natural language querying. The available literature is dominated by machine translation and other linguistic processing systems. Swahili NLQ translation algorithms would be useful in transforming the extracted phrases into meaning representations and the meaning representations into equivalent SQL statements.

3.0 CURRENT APPROACHES TO NL QUERY TRANSLATION

3.1 Major system architectures

There are four main architectural designs that are used in most systems reported in the literature. These are:

- Pattern matching systems
- Syntax-based systems
- Semantic grammar systems
- Intermediate representation languages (Interlingua approach)

3.1.1 Pattern Matching Systems

This approach uses rules that are used to process the users' request. An interesting approach to text retrieval and phrase processing (developed for the Japanese Language) can be found at [12, 22] among others. If the request contains some particular key words, the rules guide the computer response. A famous program that uses this approach is the ELIZA program. Wikipedia encyclopedia, [21], describes this system as follows:

> ***"ELIZA*** *is a famous 1966 computer program by Joseph Weizenbaum, which parodied a Rogerian therapist, largely by rephrasing many of the patient's statements as questions and posing them to the patient. Thus, for example, the response to "My head hurts" might be "Why do you say your head hurts?" The response to "My mother hates me" might be "Who else in your family hates you?"'"*

Pattern matching approach is popular because of its simplicity. In this case no elaborate parsing or interpretation modules are required. However, systems that use the pattern matching approach will sometimes fail to correctly interpret the input; making them unsuitable for real applications within the public sector.

3.1.2 Syntax-based Systems

There are two primary structural systems of language that are used to analyze natural language namely phonology and grammar. Phonology studies the sounds of a language and is most applicable to speech processing. Grammar combines morphology and syntax. Morphology is the study of the structure of a word. Syntax is the study of formal relationship between words, which is the study of sentence structure. Words are clustered into classes called parts of speech (POS). POS are important because they provide critical information about the word, its neighbors and its pronunciation, as well as extracting the word stem in information retrieval systems. Dionysius Thrax of Alexandria in 100 B.C. proposed eight categories of POS which are still in use today, namely noun, verb, pronoun, preposition, adverb, conjunction, participle and article.[10]. Other POS have since been incorporated into natural language processing.

Words relate with others to form phrases. A sentence can be understood when it has a structure that is logical and intuitive for humans. Computers use rules known as phrase structure rules to analyze form (or syntax) of sentences. For example

$$S \longrightarrow NP + VP$$

In syntax-based systems a natural language question is analyzed by parsing. The resulting parse tree is directly mapped to an expression in some database query language. Syntax-based systems use a grammar that describes the possible syntactic structures of user's questions [3]. Consider the following parse tree and a possible database mapping.

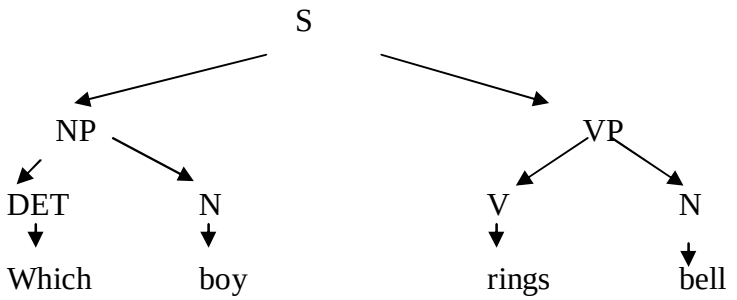


Fig. 3.1 Parse Tree and A Possible Database Mapping-Syntactic parsing Approach.

An analysis on Natural Language in Databases (NLIDB) architectures that builds on previous studies by the Russian Research Institute of Artificial Intelligence [8] reveals that most systems until the mid 1980s adopted this approach. Some popular systems included Chat-80, Language Access and Masque among others [4, 8]. The motivation behind this approach was the belief that mapping of syntactically parsed results on to a database is a tractable problem.

However, this approach is brittle because a question can take many different forms including grammatically incorrect yet semantically sensible ones. In addition, the systems developed are not portable; they work only for the specific database for which they were developed. As a result most current systems based on this approach are research laboratory prototypes that have not been ported to practical real world domains. The database profession often dismisses them as exotic demonstrations which are impracticable for database interfaces. [8]

3.1.3 Semantic Grammar Systems

In this paradigm input questions are parsed and the resulting semantic tree is mapped directly onto a database query. In semantic grammar systems the parse tree does not correspond to syntactic rules but to special grammatical categories that are carefully selected to enforce semantic constraints. The grammar categories are also chosen to facilitate mapping to the database. The following example adopted from [2] helps illustrate this concept.

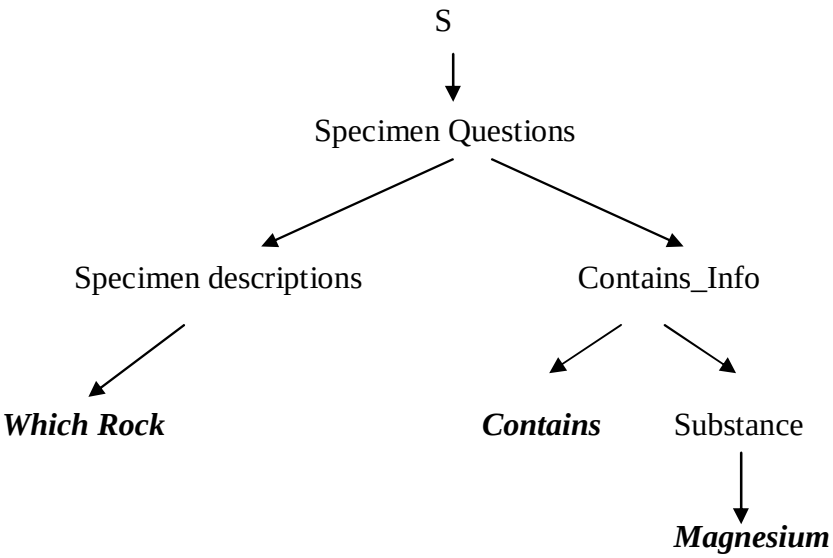


Fig. 3.2 Parse Tree and A Possible Database Mapping- Semantic Grammar Approach

In the above example, the type of specimen question directs the system to consult the database tables containing information about specimens rather than other types of information. It also directs the system to tables containing specimen contents such as a 'contains_info' table. One such system reviewed is the *INBASE* database access system [8].

Like systems based on the syntactic approach, semantic systems are rarely portable because of the amount of domain knowledge that is required to build an ontology to underpin the semantic grammar. A new semantic grammar has to be written whenever the NLIDB is configured for a new knowledge domain. However this approach is active in the research literature and has been cited in several research works such as [6].

3.1.4 Intermediate Representation Languages (Interlingua Approach)

In this approach the natural language question is first transformed into an intermediate logical query expressed in some internal meaning representation language. The representations are independent of the database structure. Schematically this would be as shown here below

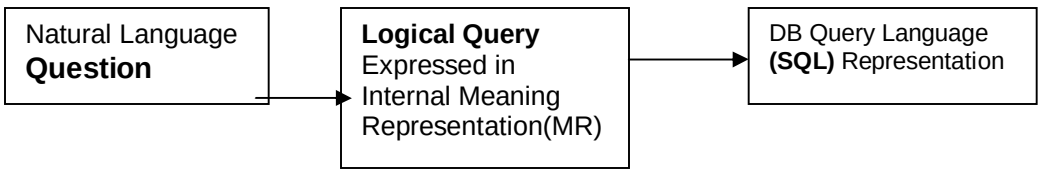


Fig. 3.3 Stages of Processing in Interlingua Approach

This approach has been cited in many current research works as a preferred approach for natural language query processing as well as other NL understanding problems such as machine translation, information retrieval and construction of dictionaries [2, 7, 11, 17, 23]. An interesting example of a system using this approach is the TAMIC-P research

project [1]. The input queries are parsed and mapped onto a representation in quasi-logical form which serves as a basis for the required database access.

4.0 SYSTEM DESIGN

We have selected a system design that builds upon the best practices in the light of the strengths and weaknesses of current approaches as described in the literature as reviewed in the previous. A system that provides an efficient, cost-effective and successful interface between the general public and government information must be capable of being practically applied in a real world domain and therefore should be:

§ User-friendly,
§ Robust,
§ Portable, and
§ Accurate.

This research proposes the use of Interlingua approach but with some modifications. The semantic processing component will be emphasized while the syntactic processing will be limited to phrase recognition.. This approach does not require a deep morphological and syntactic analysis which sacrifices robustness and flexibility. Frames will be used in processing meaning representations. These are occasionally referred to as 'Attribute-Value trees/tables' in some reviewed works, see [9] among others. This modified Interlingua approach is comparable to that used by [8, 9, 16]. These references have reported good results for Russian, Korean and Spanish languages respectively.

4.1 PROPOSED SYSTEM FRAMEWORK

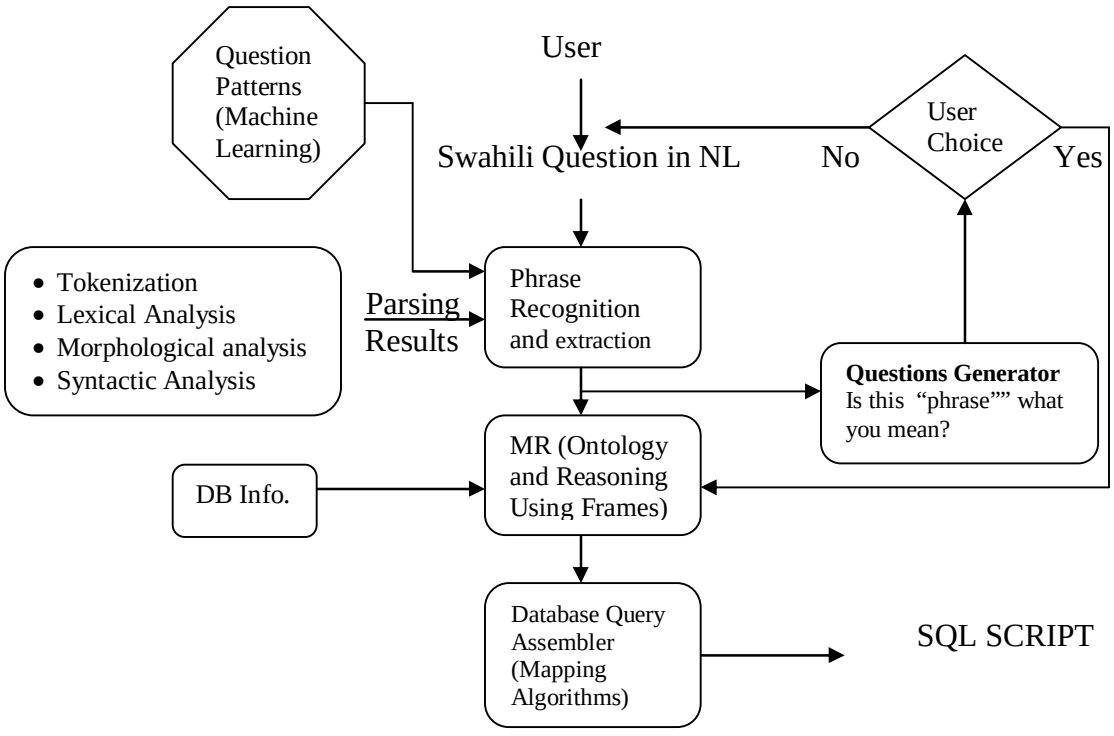


Fig. 4.1 Framework for Conversion of NL Text to SQL Text Output

The above structure shows the basic steps required to convert NL text into an SQL text. Well documented Swahili NL processors, such as reported in [4], for tokenization, lexical analysis, morphological analysis and syntactic analysis can be used in the above framework. The framework incorporates an interactive question repair module which will enhance accuracy of the output. Machine learning module studies the pattern of the input questions and provides for improved efficiency in the mapping of NL inputs to actual meaning. The efficiency improves with usage of the system.

4.2 An Example Illustrating Framework's Working Principles

Suppose the following query is to be converted into the equivalent SQL statement

Swahili: *"Nipe orodha ya wanawake wote wenye vyeo vya juu"*
English: *'Give me a list of salaries for all women with plump jobs'*

The statement is first broken down into the various logically cohesive tokens (lexeme). All the words are morphologically analyzed to give the stem and other POS information. Thus

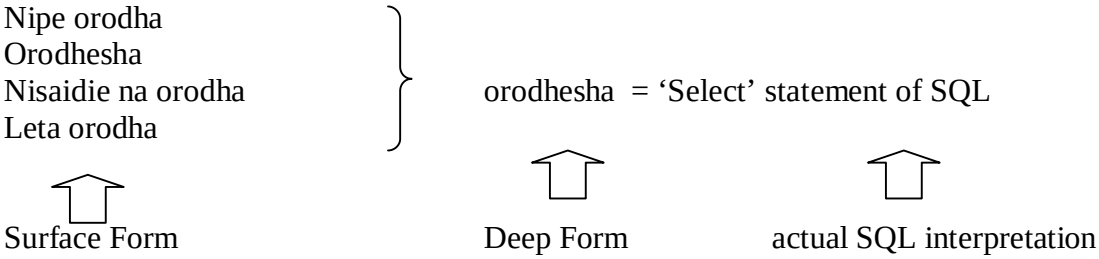
Nipe >> is a verb and has two morphemes ni- and pe. etc.

The sentence is then syntactically analyzed to get the various parse trees. From these parse results useful information on word neighbors is obtained. This information is necessary in the recognition of noun and verb phrases which are used in the MR ontology
In the above example we obtain

Swahili NL: Nipe orodha >>ya>> ***mishahara*** ya ***wanawake*** wote>> wenye >>***vyeo vya juu***
English equivalent: Give me a list >>of >>***salaries*** for all ***women*** >>with >>***plump jobs***
Thus

```
SELECT Salary
FROM Worker_Table
WHERE Gender = 'f' AND Grade = 'senior'
```

'Nipe orodha' is a verb phrase that maps onto SELECT
'Mishahara' is Noun appearing immediately after the verb phrase and provides a hint to the 'TABLE NAME' and the 'COLUMN' to go along with the SELECT statement..
The rest of the statement provides a filtering criteria for WHERE clause.
The framework provides for synonyms by mapping surface form of statement to deep form. For example



A full discussion of the actual ontologies and mapping algorithms are beyond the scope of this paper.

5.0 Discussion

We have presented a framework which can effectively parse Swahili NL input text to its corresponding SQL output. We have stated how efficiency and accuracy are guaranteed. Here we discuss how the framework achieves the desired levels of comprehension of ungrammatical sentences, portability and robustness.

.

5.1 Comprehension of Ungrammatical Sentences

As noted above human beings do not necessarily speak grammatically correct NL. In a region like East Africa this problem is compounded by the fact that Swahili speakers are highly affected in the way they speak and write Swahili by their first languages. This paper proposes a phrase-based processing as opposed to full syntactic processing. Phrases are identified using some standard context free grammar (CFG) rules and their extensions. These extensions will be a new feature and are still being studied. The use of meaning representations and identification algorithms enhances recognition of a wide array of synonymous phrases whether grammatically accurate or not.

For example mimi ***iko taka*** maji ……. Me want water

Na***taka*** maji ……..I want water

The phrase ***iko taka*** should map onto the key word ***taka***

This means that there will be extension of the standard phrase structure grammars, noted in section 2.1.1 above, to accommodate commonly used phrases though grammatically inaccurate

5.2 Portability

The designed frame envisages modules independence. This design accords a better portability. The methodology for phrase recognition is universal to all domains. However the lexicon within the MR ontology is domain specific and has to be changed whenever a new domain is selected. This is not unique to this approach and form based applications also experience the same; hence it is not a major concern.

5.3 Robustness

The designed model provides for robustness from various considerations. The user has an opportunity to repair the query and hence provide for some degree of reliability. The use of machine learning further improves this aspect by providing previously successful results. With an anticipated high number of users querying government data repositories the module is expected to increase efficiency and accuracy hence robustness.

5.4 Adoption Of The Proposed Framework To Other Languages

The proposed framework is language independent and may be adapted for virtually any natural language. Language specific information is incorporated within phrase recognition and extraction module only. The same MR ontologies and database query assembler can be used for any language. Further machine learning and query repair modules are both language and domain independent.

5.5 Limitations Of the Framework

Translation content within the Meaning representation framework changes with domain and specific database architecture. This makes the system less portable as would be required. Further works may be required to improve portability.

The applicability of this model would further be enhanced by researching into the area of Swahili speech processing with an aim of incorporating a speech input and output system**.**

CONCLUSION

This paper has brought forth some key outstanding research issues in Swahili NLQ processing. They include lack of computerized phrase identification mechanisms for Swahili texts, difficulties in computerized information extraction using traditional syntactic processing methods and the lack of algorithms that translate Swahili NL queries into Structured Query Language (SQL). The paper has presented a framework that can achieve user-friendly, accurate, efficient, robust and generally portable interface between the general public and government information. The framework can be adopted for any language or domain.

.

REFERENCES

1. Alexandra K., Johannes M. and Herald T., (1998). The treatment of Noun phrase Queries in a Natural Language database Access System - German NL access to Austrian Social Insurance Institution database for Farmers. SiteSeer Database. WWW.SiteSeer.ist.psu.edu

2. Androutsopoulos, Ritchie G.D., Thanisch P., (1995). Natural Language Interfaces to Databases- an Introduction. SiteSeer Database WWW.SiteSeer.ist.psu.edu

3. Androutsopoulos,. Ritchie G., Thanisch, P. (1993) MASQUE/SQL An Efficient and Portable Natural Language Query Interface for Relational Databases. SiteSeer Database, http://citeseer.ist.psu.edu/androutsopoulos93masquesql.html

4. Arvi, H., (1999) Swahili Language Manager- SALAMA. Nordic Journal of African Studies Vol. 8(2): 139-157. http://www.njas.helsinki.fi/pdf-files/vol8num2/hurskainen.pdf

5. Cabinet,2004. E-Government strategy -The strategic Framework, Administritive Structure, Training Requirements and Standardization Framework. Cabinet Office, Republic of Kenya. Government Printer, Nairobi Kenya.

6. Genereux M., (2005). Efficient Semantic Parsing Of Conversational Speech. Proceeds of 2nd Baltic Conference on Human Language Technologies, April 4-5, 2005, www.ioc.ee

7. Hurskainen A., Sewangi.S, and Ng'ang'a W.,(2006)SALAMA - Swahili Language Manager. Center for Scientific Computing, http://www.njas.helsinki.fi/salama/

8. Inbase System,(2002). Russian Research Institute of Artificial Intelligence, www.inbase.artint.ru

9. Jung, H. and Lee. G., (2002). Multilingual question answering with high portability on relational databases. Proceedings of the 2002 conference on multilingual summarization and question answering - Volume 19. Pages: 1 – 8. Published by Association for Computational Linguistics, Morristown, NJ, USA. http://portal.acm.org/citation.cfm?id=1118847&dl=ACM&coll=&CFID=15151515&CFTOKEN=6184618

10. Jurafsky D., and Martin J.H., (2003). Speech and Language Processing-An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson ed. Series, New Delhi, India.

11. Karttunen L.,(1989). Translating from English to Logic in Tarski's World. Journal of Information Sience and Engineering, Vol. 5 No. 4, pp.323-348. www.iis.sinica.edu.tw/JISE

12. Kitani T., Eriguchi Y. and Hara M. (1994). Pattern Matching and Discourse Processing in Information Extraction from Japanese Text. AI Access Foundation and Morgan Kaufmann Publishers. www.cs.washington.edu/research/jair/contents/complete-with-abstract.html

13. Miriti E. (2006). Swahili Dictation System (Speech to Text Machine Translation). An Unpublished Thesis for Master of Science(Applied Computing), School of Computing and Informatics, University of Nairobi

14. Ng'ang'a W.J., (2005) Word Sense Disambinguation of Swahili-Extending Swahili Language Technology with Machine Learning. PhD Thesis, University of Helsinki, Finland. http://ethesis.helsinki.fi/julkaisut/hum/aasia/vk/nganga/wordsens.pdf

15. Ohly,R.1982. " Report on Lexicographic Research at the Frendship Textile Mill." In Kiswahili. Journal of the Institute of Kiswahili Research. Dar es Salaam. University of Dar es Salam,73-86

16. Rodolfo, A., Gelbukh A., Gonzalez, J., Ruiz, E., Mejia, M., and Sanchez, P (2002). Spanish Natural Language Interface for a Relational Database Querying System. Proceedings of the 5th Text, Speech and Dialog International Conference (Czech Rep.), pp. 123. Published by Springer Berlin, Heidelberg, Germany.

17. Rojas J., Torres J. A Survey in Natural Language Databases Interfaces, Proceedings of 8th International Congress on Computer Science Research. Instituto Tecnologico de Colima, Colima, Mexico, Nov. 2001, pp 63-70.

18. Sewangi S., (2002) Computer- Assisted Extraction of Terms in Specific Domains- The Case of Swahili. PhD Thesis, University of Helsinki Finland. http://ethesis.helsinki.fi/julkaisut/hum/aasia/vk/sewangi/computer.pdf#search=%22sewangi%20phd%20thesis%22

19. Temu, C.W. 1984. " Kiswahili Terminology: Principles Adopted for the enrichment of the Kiswahili Language." In Kiswahili. *Journal of the Institute of Kiswahili Research.* Dar es Salaam. University of Dar es Salam,112-127-86

20. Wamitila K.W., (2005). "*Chemichemi za Kiswahili*." Published by Longhorn publishers , Nairobi Kenya.

21. Wikipedia., (2006). http://en.wikipedia.org/wiki/ELIZA

22. Yoshioka M., Kageura K., Kando N. and Keizo O. (1998) Phrase Processing Methods for Japanese Text Retrieval. SiteSeer Database. WWW.SiteSeer.ist.psu.edu

23. Zarate J., Rodolfo, A., Gelbukh A., Padron I.(2003). A portable Natural Language Interface for Diverse Databases Using Ontologies. Proceeds of CICLing International Conferences, Mexico. Pdf file- http://www.gelbukh.com/CV/Publications/2003/CICLing-2003-DB.pdf

24. Androutsopoulos I., A Principled Framework for Constructing Natural Language Interfaces to Temporal Databases. PhD Thesis, University of Edinburgh. 1996

25. Ginter G., *et al*. Information Extraction From Biomedical Text: The Biotext Project. Proceeds of Second Baltic Conference on Human Language Technologies. 2005. www.ioc.ee/hlt2005/hlt2005.pdf

26. Chae Jinseok and Lee Sukho (1995). Natural Language Query Processing in Korean Interface for Object-Oriented databases. SiteSeer Database. WWW.SiteSeer.ist.psu.edu