

# Semantic Interoperability and Lexicon Development

Steven Lulich\*, Paul Thompson†

\* Program in Linguistics  
& Cognitive Science  
Dartmouth College  
Hanover, NH 03755  
[steven.m.lulich@dartmouth.edu](mailto:steven.m.lulich@dartmouth.edu)

† Institute for Security Technology Studies  
Dartmouth College  
45 Lyme Road, Suite 200  
Hanover, NH 03755, U.S.A  
[Paul.Thompson@dartmouth.edu](mailto:Paul.Thompson@dartmouth.edu)

## Abstract

Much recent question answering research has focussed on supporting the textual retrieval needs of intelligence analysts. Question answering may also play a role in other less textual domains, such as sensor networks, or the Joint Battlespace Infosphere (JBI). We propose a connectivistic database to serve as the core of a lexicon which may be used to improve current methods of question answering, as well as other natural language and ontology processing application

## 1. Introduction

The question answering vision (Carbonell et al., 2000) and roadmap (Burger et al., 2000) documents describe a five year program for research and development for question answering systems with a focus on how such systems could support the needs of an intelligence analyst. DARPA's Office of Information Exploitation (IXO) program has the mission to ". . . develop sensor and information systems with application to battle space awareness, targeting, command and control, and the supporting infrastructure required to address land-based threats in a dynamic, closed-loop process." IXO is developing 1-, 5-, and 20-year vision statements to meet the challenges of these systems. These dynamic information environments require intelligent middleware to broker services to connect information users and sources. For example, users pose natural language questions, which must be translated into the query languages and ontologies of the heterogeneous systems making up the JBI (United States Air Force Scientific Advisory Board, 1999, 2000; Infospherics, 2001). While technologies in this area will build on current DARPA programs providing tools for efficient human creation of ontologies (DARPA Agent Markup Language, 2002; DARPA Rapid Knowledge Formation, 2002), because of the dynamic, rapidly changing environment represented by the JBI, it is necessary that more automated approaches to semantic interoperability be developed, as well.

We suggest the desirability of a connectivistic database to serve as the core of a lexicon which may be used to improve current methods of question answering, as well as other natural language and ontology processing applications. Specifically, we illustrate the use of such a lexicon in the Joint Battlespace Infosphere (JBI). Related work has been done on statistical tools that automate the process of mapping from one ontology or grammar to another (Thompson, 2001). We are interested in building on this work, as well as using mixed-initiative approaches (Haller et al., 1999) to provide human input, where needed.

## 2. Lexicon Development: Application of Linguistic Knowledge to Natural Language Processing

## 2.1. Properties of natural language which may be mimicked computationally

Three aspects of natural language are submitted for consideration:

- Grammars consist of categories which may be cognitively manipulated synchronically or altered diachronically (Heine, 1997), such as phones, morphs, words, and grammatical classes. The categories within grammars are defined with respect to each other, much as the words of a dictionary are defined with respect to other words in the dictionary, and do not therefore line up evenly across languages (Whaley, 1997). For instance, the study of languages as diverse as English, Tagalog, Manchu, and !Xhosa has resulted in the understanding that lexical classes in different languages do not all conform to the same mould. Some languages employ lexical classes which are not employed in English, and vice versa. Furthermore, the same class in different languages may not be easily reconciled with each other, and the distinctions between classes, even noun and verb, can sometimes become blurred. Morphologists and psycholinguists such as Joan Bybee (1988) and Ardi Roelofs (1992), to name only two, have explored the idea of a connectivistic lexicon with some success, both conceptually and experimentally.
- Grammars do not consist only of minimal units and rules for combining them. It has been found that the human brain stores a far more redundant amount of linguistic information than had previously been thought. Work with aphasic patients shows that the use of rules in combining morphemes may be thought of as a back-up method for producing morphologically complex words when access to the lexicon fails (Badecker & Caramazza, 1998). Psycholinguistic experiments have shown that the timing of lexical access for morphologically simple words is not significantly different from the timing of lexical access for morphologically complex words, and phonetic and psycholinguistic studies indicate that some prosodic structures are stored as whole units alongside of the individual segments of which they are comprised (Levelt, 1999; Grzegorz Dogil, personal communication).
- Grammars are learned best by immature brains – brains with degraded short term memory – which may learn only general principles of grammar before narrowing down to specific principles (Deacon, 1997). Deacon outlines work done by others in cognitive and computer science which involved training of neural networks to learn a grammar to a relatively large degree of accuracy when the “short-term memory” of the network was disturbed. Studies by MacWhinney (1978) and Peters (1983) indicate that generalizations (rules) gradually emerge from stored rote forms, which are initially processed and stored as unanalyzed wholes, cf. (Bybee, 1988). These studies corroborate both the work done by Deacon, and the evidence that linguistic data stored in the lexicon is often redundant.

## 2.2. Proposal for the design of a lexicon which mimics these properties

A lexicon with five main components may serve to mimic these properties of natural language: a Pattern Finding Engine (PFE), Short Term Memory (STM), Long-Term Memory (LTM), Connectivistic Database (CD), and an Anchor Set (AS),

### 2.2.1. Pattern Finding Engine and Memory

The Pattern Finding Engine (PFE) searches a text for patterns, and, during the training phase of the lexicon, stores those patterns in the Short-Term Memory (STM), while the strings predictable from those patterns are stored in the Connectivistic Database. For instance, starting from scratch, the PFE recognizes a sentence such as “Johnny ate the apple” as a single unit. This imitates the theory derived from the work of Deacon, MacWhinney, and Peters above. This single unit is stored as a whole in the CD as an object of class “lexical unit.” Exposure to more sentences, such as “Johnny ran away” and “The apple is red”, enables the PFE to recognize “Johnny” and “the apple” as units, and to store them in the CD, along with “ran away” and “is red”. Further exposure to sentences such as “Apples taste good” and “Jack and Jill ran up the hill” allows the PFE to recognize “ran” as separate from “away again”, and “s” as a morpheme attached to “apple”.

Initially, PFE is not better than chance at finding correct patterns. Therefore, potential patterns are stored in STM. As more and more occurrences of patterns in STM are found by PFE, the patterns in STM are stored in Long-Term Memory (LTM). Because some units larger than the segment or the word may occur with great frequency, the work of PFE together with STM and LTM allows an imitation of the theory that the lexicon is not redundancy free. This also allows us to capture idioms as whole chunks (Nunberg et al, 1994).

### **2.2.2. Connectivistic Database**

An object of class “lexical unit” represents all of the information concerning a single unit. Within the object of class “lexical unit” is a set of objects of class “link”. Each object of class “link” contains two variables: a pointer, pointing to one other object of class “lexical unit”; and a value corresponding to the strength of that connection. Each “lexical unit” also contains an activation value, which records and keeps track of the activation of that unit at all times. Activation is a measure of the probability that a certain unit will be the next one chosen out of the lexicon, and is determined by the amount of activation flowing to it through its connections with other activated units. Each “lexical unit” also has an abstract position variable, represented by an n-dimensional vector, which identifies a location for the “lexical unit” in an abstract n-dimensional Minkowsky space.

Throughout the training phase, with the help of PFE, STM, and LTM, the CD automatically organizes itself into an n-dimensional Minkowsky space. Categories are automatically approximated by defining opposing categories with respect to each other along a similar dimension. Sets of categories which are not defined with respect to each other are defined along different dimensions. Such definitions may be approximated without prior human or machine coding (Klein, 1998; Levine et al., 2001).

### **2.2.3. Anchor Set**

Initial training of the lexicon is supervised by a human assigning certain “lexical units” to corresponding absolute concepts. Such “anchor points” provide the basis for translation from one grammar or ontology to another via the lexicon. English “chair” and German “Stuhl”, for instance, refer to roughly the same concept. Therefore, the word “chair” in an English trained lexicon, and the word “Stuhl” in a German trained lexicon will both be anchored to the concept of “CHAIR”. The Anchor Set (AS) can be used then to manipulate and align the abstract n-dimensional vector spaces of the two lexicons such that, by extrapolation, lexical units with nearly identical position vectors should theoretically be nearly identical in meaning or use, depending on the dimension. The more anchor points that are explicitly taught to the AS, the more accurate this alignment will be.

## **2.3. Discussion**

To the best of our knowledge, though the ideas and evidence outlined in this paper in favor of a connectivistic view of the lexicon have been explored by linguists already, there has been no attempt to apply such a model to challenges in natural language processing. Certainly this may partially be attributed to the fact that the computing power necessary to undertake such a task has not long been available.

We believe that development of such a lexicon is relevant to Question Answering technology in several ways. First, the lexicon, whatever shape it may take, is an important and central part of any natural language processing application. Without it, language is simply noise. We believe therefore that the form of the lexicon has a direct effect on the overall performance of the application. Second, in answering a single question, it is often necessary to extract information from multiple sources of varying media and ontologies. The information coming from these disparate sources must somehow be fused together and outputted into yet another ontology or medium. Because this conception of a lexicon is easily trained, it is easily transportable across multiple domains and ontologies or grammars. As discussed in section 2.2.3, the Anchor Set allows translation from one ontology to another via the lexicon, thus enabling this kind of fusion of information. Finally, though certainly not exhaustively, the automatic categorization of words along different dimensions, and the connections between words may be helpful as a tool for word sense disambiguation.

### 3. Questions in the Infosphere

#### 3.1. Background

Question answering in heterogeneous sensor networks involves some of the same issues as question answering in more textual domains, but also introduces other aspects. The answer to the question may not exist in the network at the time the question is asked. Sensors may need to be tasked to provide the answer. A mapping must be made between the language of the user and the descriptions of the functionalities of various sensors. There is high transaction volume in the Joint Battlespace Infosphere (JBI) and questions may overlap in various ways. Efficient question answering calls for query planning and optimization along the lines of work done in relational databases (Jarke & Koch, 1984) and knowledge bases, but with additional factors introduced by the distributed, mobile, highly dynamic nature of sensor networks. Also, because much of the data in these networks will be structured, question answering in this environment can also build on research on natural language interfaces to relational databases (Adroutsopoulos et al., 1995; Urro & Winiwarter, 2001).

The JBI consists of client users, databases, sensors, and filtering or fusion operations. These filtering or fusion operations are carried out by fuselets, lightweight data fusion elements. Fuselets use simple logical rules to take inputs from other elements of the JBI, such as sensors, or other fuselets, to derive fused information. The functionality of each fuselet is described using a Fuselet Markup Language (FML). The JBI is implemented as a publish and subscribe architecture, where each fuselet publishes its services and subscribes to the outputs of other elements of the JBI. Questions in the JBI are answered by breaking the question into components and efficiently routing the components through the JBI network of fuselets, databases, and sensors.

Although ontologies may be provided for various sub-domains, it may be necessary to rapidly create and map among ontologies on the fly. For example, a fuel truck may be represented in separate ontologies for target tracking and for logistics. It must be possible to: a) determine that the two representations are of the same type of entity, b) reason within the joint probability space represented by the two ontologies, and c) answer questions by fusing information from the two domains. We will investigate a variety of tools to achieve semantic interoperability. In addition to the linguistic approaches to lexicon development discussed in section 2, we plan to explore statistical, text-based mapping and subsumption tools (Woods, 1997; Buckland et al., 1999; Gey et al., 2001; Schatz, 2002).

#### 3.2. A JBI Fuselet Example

As a simplified example of question answering in the Infosphere, consider the following. In a battlefield situation when an enemy target is to be fired upon, it is first necessary to ascertain that no friendly assets are in the vicinity that might be adversely affected. A subset of the JBI involving a network of sensors, radio transmitters operated by groups of soldiers, advanced Land Warrior personal GPS systems, current roster information, other sources of information, and fuselets would be needed to make this determination. The current location, velocity, and vector of all friendly assets would need to be determined. If processing this information takes too much time, the target opportunity might be missed. If the enemy target is fired upon without the information being processed accurately, friendly assets may become casualties. Personnel in the tactical operation center would submit a natural language query, "Are any friendly assets in danger of being hit, if the target at UTM grid coordinate XY123456 is fired upon?" This query would then be interpreted by the question answering system. Fuselet 1 would aggregate the outputs from the soldiers' radio transmitters. Fuselet 2 would aggregate the output of the GPS systems. Fuselet 3, with situational tracking software, would fuse the outputs of Fuselets 1 and 2. Fuselet 4 in the personnel services center would fuse outputs from databases with current roster information, as well as with outputs from other databases making adjustments to the current roster, e.g., lists of soldiers on medical leave. Fuselet 5 would fuse the outputs of Fuselets 3 and 4 and produce as output a report for the tactical operations center, answering the query.

## 4. Conclusions

We intend to address question answering issues in the JBI, in particular those concerning closed-loop sensor networks. Our domain has some overlap with that of the intelligence analyst described in the question answering vision and roadmap documents, but has significant differences, as well. We intend to build a sensor network integrated with textual messages. We will make use of ontologies, such as a sensor markup language, but we will also explore connectivistic lexicon, corpora linguistic, and other techniques to learn about our domains in a more dynamic manner, as necessary.

## 5. References

- Androutopoulos, I.; Ritchie, G.D.; & Thanisch, P. (1995). Natural language interfaces to databases – An introduction. *Journal of Natural Language Engineering*, 1(1), p.29-81
- Badecker, W. & Caramazza, A. (1998). Morphology and Aphasia. In A. Spencer, & A.M. Zwicky (Eds.), *The Handbook of Morphology* (pp. 390-405). Oxford: Blackwell Publishers Ltd.
- Buckland, M., Chen, A., Chen, H., Gey, F., Kim, Y., Lam, B., Larson, R., Norgard, B., & Purat, Y. (1999). Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies. *D-Lib Magazine*, 5(1).
- Burger, J.; Cardie, C.; Chaudhri, V.; Gaizauskas, R.; Harabagiu, S.; Israel, D.; Jacquemin, C.; Lin, C.; Maiorano, S.; Miller, G.; Moldovan, D.; Ogden, B.; Prager, J.; Riloff, E.; Singhal, A.; Shrihari, R.; Strzalkowski, T.; Voorhees, E.; Weischedel, R. (2000). Issues, tasks and program structures to roadmap research in question & answering (q&a). Gaithersburg: National Institute of Standards and Technology.
- Bybee, J. (1988). Morphology and Lexical Organization. In M. Hammond & M. Noonan (Eds.), *Theoretical Morphology: Approaches in Modern Linguistics* (pp. 119-142). San Diego, CA: Academic Press.
- Carbonell, J.; Harman, D.; Hovy, E.; Maiorano, S.; Prange, J.; & Sparck Jones, K. (2000). Vision statement to guide research in question & answering (Q&A) and text summarization. Final version 1. Gaithersburg : National Institute of Standards and Technology.
- DARPA Agent Markup Language (DAML) (2002). <http://dtsn.darpa.mil/ixo/daml%2Easp>.
- DARPA Rapid Knowledge Formation (RKF). (2002). <http://dtsn.darpa.mil/ixo/rkf%2Easp>.
- Deacon, T. (1997). *The Symbolic Species: The Co-evolution of Language and the Brain*. New York: W.W. Norton.
- Gey, F.; Buckland, M.; Chen, A.; & Larson, R. (2001). Entry vocabulary – a technology to enhance digital search. In *Proceedings of HLT 2001: First International Conference on Human Language Technology Research* (pp. 91-95). San Francisco: Morgan Kaufmann.
- Haller, S.; McRoy, S.; and Kobsa, A. (Eds.). (1999). *Computational Models of Mixed-Initiative Interaction* Boston: Kluwer.
- Heine, B. (1997). *Cognitive Foundations of Grammar*. New York: Oxford University Press.
- Infospherics: Science for Building Large-scale Global Information Systems. (2001). <http://actcomm.dartmouth.edu/infospherics/>
- Jarke, M. & Koch, J. (1984). Query optimization in database systems. *Computing Surveys*, 16(2), 111--152.
- Klein, A. (1998). *Textual Analysis Without Coding: It Can be Done*. Dissertation, Mathematical Social Sciences, Dartmouth College.
- Levelt, W.J.M. (1999). Producing spoken language: a blueprint of the speaker. In P. Hagoort & C.M. Brown (Eds.) *The neurocognition of language* (pp. 94-122), Oxford: Oxford University Press.
- Levine, J.H.; Klein, A.; & Mathews, J. (2001). Data Without Variables. *Journal of Mathematical Sociology*, 23(3), 225--273.
- MacWhinney, B. (1978). *The Acquisition of Morphophonology*. Child Development Publication, Chicago: University of Chicago Press.
- Nunberg, G; Sag, I.; & Wasow, T. (1994). Idioms. *Language*, 70, 491--538.
- Peters, A.M. (1983). *The Units of Language Acquisition*. Cambridge, U.K.: Cambridge University Press.
- Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, 42, 107--142.
- Schatz, B.R. (2002). The Interspace: Concept navigation across distributed communities. *IEEE Computer*. 35(1), 54--62.

Thompson, P. (2001). Classification Crosswalks: From Interchange to Interoperability. Classification Crosswalks: Bringing Communities Together The 4th NKOS Workshop at ACM-IEEE Joint Conference on Digital Libraries (JCDL).

United States Air Force Scientific Advisory Board. (2000). Report on Building the Joint Battlespace Infosphere, vol. 1 Summary SAB-TR-99-02.

United States Air Force Scientific Advisory Board. (1999). Report on Building the Joint Battlespace Infosphere, vol. 2 Interactive Information Technologies SAB-TR-99-02.

Urro, R. & Winiwarter, W. (2001). Specifying Ontologies – Linguistic Aspects in Problem-Driven Knowledge Engineering. In Proceedings of the 2nd International Conference on Web Information Systems Engineering, Los Alamitos, IEEE Computer Society Press.

Whaley, L.J. (1997). Introduction to Typology: The Unity and Diversity of Language. Thousand Oaks, CA: Sage Publications.

Woods, W.A. (1997). Conceptual indexing: A better way to organize knowledge. Sun Microsystems Research Technical Report TR-97-61.