# Finding Genealogy Facts with Linguistic Analysis

Peter Lindes, Deryle W. Lonsdale, David W. Embley
Brigham Young University

## Abstract

OntoSoar is a system to extract information from genealogy texts and populate a conceptual model with the extracted information. It does this by doing a deep linguistic analysis and then mapping the meaning structures found to an ontology provided by the user. The system is built using a processing pipeline that includes the open-source Link Grammar Parser using a grammar customized for this application and an innovative semantic analyzer built on the Soar cognitive architecture. Here we describe the how OntoSoar, currently still in development, works.

## Introduction

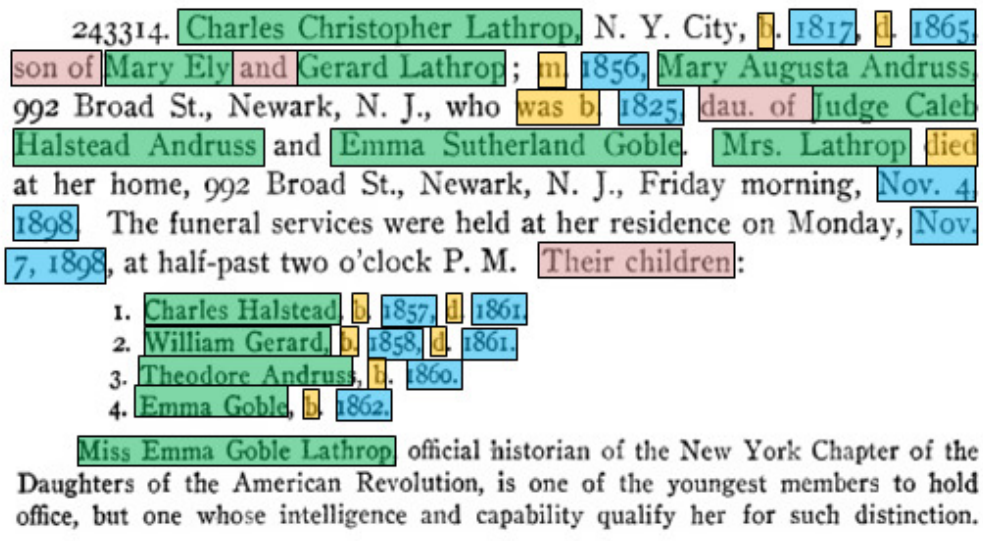
One potential large source of genealogical information is the thousands of historical books on family histories that have now been scanned and OCR'd. A typical page contains dozens of facts about people, their names, their life events, and their family relationships. A collection of 100,000+ such books of several hundred pages each has many millions of facts, but it would take an enormous amount of work to extract all this information manually from the digital texts.

Previous work has been done on automating the extraction of this information. Cimiano (2006) and Wong (2012) survey the field of extraction from text in general. Embley et al. (2011) discusses a system called OntoES that attacks the problem using a conceptual model and "extraction ontologies" that use regular expressions to find textual patterns that contain facts. Lonsdale et al. (2001) describes a system based on LG-Soar that uses natural language processing techniques and the Soar cognitive architecture to find facts in the text.

The research described here extends the work of both Embley and Lonsdale by building a more complete and robust system, somewhat like Lonsdale's LG-Soar system as adapted for genealogy texts, and integrating that with tools from Embley's OntoES system. We call the new system OntoSoar.

## Sample Texts

Figure 1 shows an example of part of a page from one of these family history books:



243314. Charles Christopher Lathrop, N. Y. City, b. 1817, d. 1865, son of Mary Ely and Gerard Lathrop; m. 1856, Mary Augusta Andruss, 992 Broad St., Newark, N. J., who was b. 1825, dau. of Judge Caleb Halstead Andruss and Emma Sutherland Goble. Mrs. Lathrop died at her home, 992 Broad St., Newark, N. J., Friday morning, Nov. 4, 1898. The funeral services were held at her residence on Monday, Nov. 7, 1898, at half-past two o'clock P. M. Their children:

1. Charles Halstead, b. 1857, d. 1861.
2. William Gerard, b. 1858, d. 1861.
3. Theodore Andruss, b. 1860.
4. Emma Goble, b. 1862.

Miss Emma Goble Lathrop, official historian of the New York Chapter of the Daughters of the American Revolution, is one of the youngest members to hold office, but one whose intelligence and capability qualify her for such distinction.

Figure 1: Sample 1 of Genealogy Text

The sample in Figure 1 shows somewhat structured text, especially the list of children. Green rectangles mark the names of people, blue are dates, yellow are event verbs, and light red are family relationship phrases. These represent the basic facts we would like to extract, and the OntoES extraction ontologies can work reasonably well on cases like this. However, many other books have much less structured, free-flowing text as shown in Figure 2:

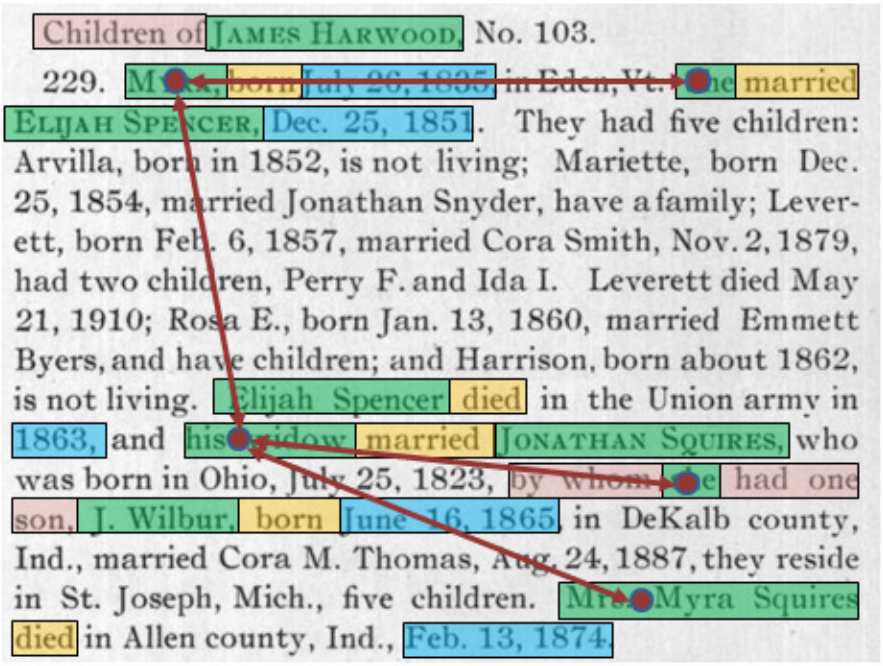



Figure 2: Sample 2 of Genealogy Text

This sample is much harder to analyze. We have just shown a few of the items that can be found here. Notice that now some of the green rectangles are not names at all but other linguistic expressions that refer to people already identified elsewhere in the text. The red dots and the arrows connecting them overlaid over the marked phrases show how linguistic analysis can decode these phrases and infer connections between them. By this analysis we can conclude that the MYRA mentioned at the beginning of the paragraph, whose maiden name we can deduce is Myra Harwood, is the same person as the Mrs. Myra Squires whose death is mentioned at the end of the paragraph. There would be no way to make these connections without a deep understanding of the meaning of the language in the text.

## System Design

OntoSoar uses innovative techniques for deep linguistic analysis based on construction grammar techniques (Hoffman & Trousdale (2013), Bryant (2008)). Combining this with ideas from OntoES and LG-Soar results in a system that is able to extract facts that neither of these previous systems could. Figure 3 shows a block diagram of the resulting system.

The natural language processing part of the system segments the text into sentences and sentence fragments, parses these to produce syntax graphs called linkages, performs semantic analysis to derive schema structures representing the meaning of the text, uses inference rules to enrich these schemas, and then uses the results to populate an ontology provided by the user in the OSMX format used by OntoES. At this writing the development is well along, but not completed yet.

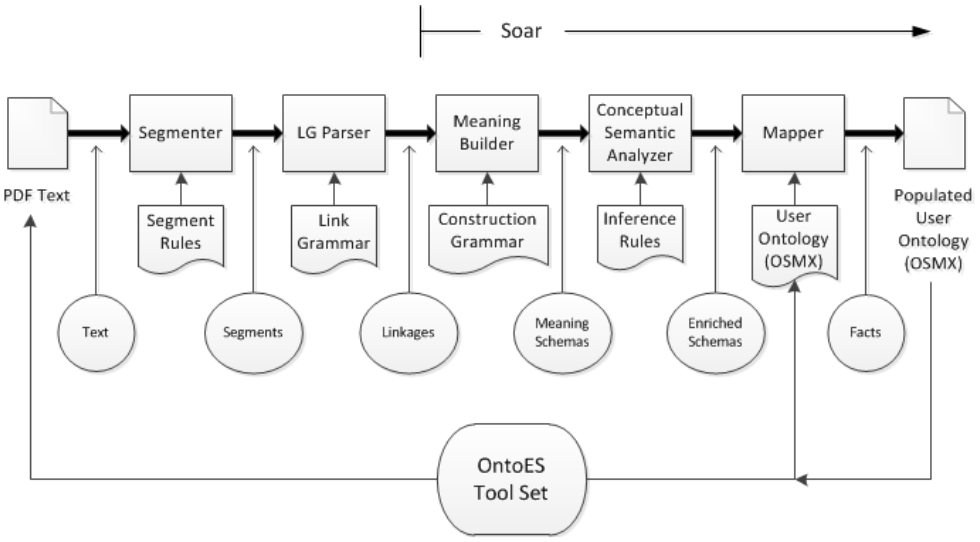



Figure 3: The OntoSoar System

The LG Parser component is built on an open-source parser called the Link Grammar Parser, whose general English grammar has been modified to adapt it to non-standard forms that are common in genealogy texts. All the semantic analysis is built from scratch in the Soar cognitive architecture, which is a powerful tool for doing this kind of analysis and inferencing. At the end of the pipeline an ontology in OSMX format built using the OntoES tools is populated with facts extracted from the text. The whole pipeline is driven by a Java program.

## How the System Works

To illustrate the logic involved in the linguistic analysis, we will look in some detail at two example sentences, one from each of the above sample texts. In Figures 4 and 5 we see an overview of the analysis of the first sentence from Sample 1, divided into two sub-cases to make the structure more visible:

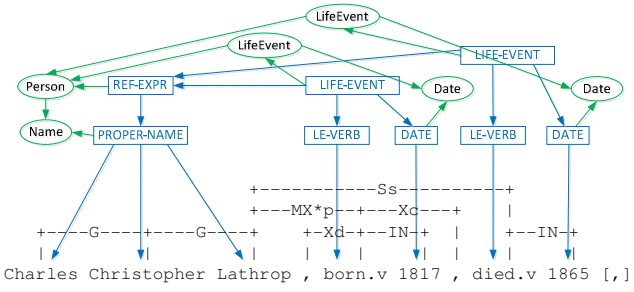



Figure 4: Linguistic Analysis – Case 1A

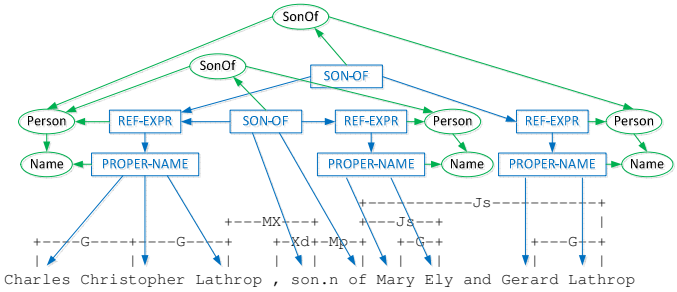



Figure 5: Linguistic Analysis – Case 1B

At the bottom of each figure we see the linkage produced by the LG Parser for this case. Using the links between words found by the parser we next apply a construction grammar approach adapted from that of Bryant (2008), producing the constructions shown in the blue rectangles. From these constructions we build meaning schemas of various types as shown by the green ovals. Figure 4 shows the analysis of the events in the subject's life, while Figure 5 shows the analysis of the family relationship part of the original sentence.

With meaning schemas derived we can collect them into a more summarized structure for additional inferencing. Figure 6 shows a complete meaning graph derived for the sentence in Case 1. In Figure 4 we see a LifeEvent schema based on the verb "born" whose subject is the Person that has the name "Charles Christopher Lathrop" and whose date is "1817." In Figure 6 this is summarized into the birth and name slots of a Person schema. We also see additional items derived by inferencing, such as father and mother relations and a Couple schema, which broaden the range of things that can eventually be mapped into possible user ontologies.

Charles Christopher Lathrop, N. Y. City, b. 1817, d. 1865, son of Mary Ely and Gerard Lathrop ;

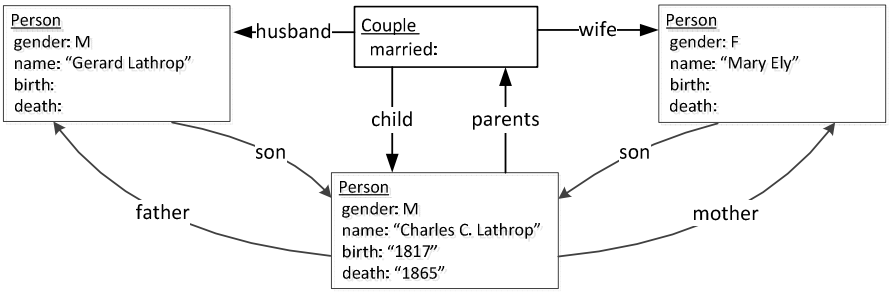
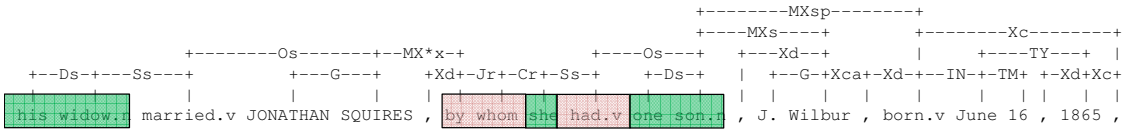



Figure 6: Meaning Summary for Case 1

Case 2 is the sentence fragment from Sample 2 starting with "his widow". For this case we show only the original sentence, the linkage produced by the LG Parser, and the resulting meaning summary in Figure 7:

his widow married JONATHAN SQUIRES, ..., by whom she had one son, J. Wilbur, born June 16, 1865, ... .



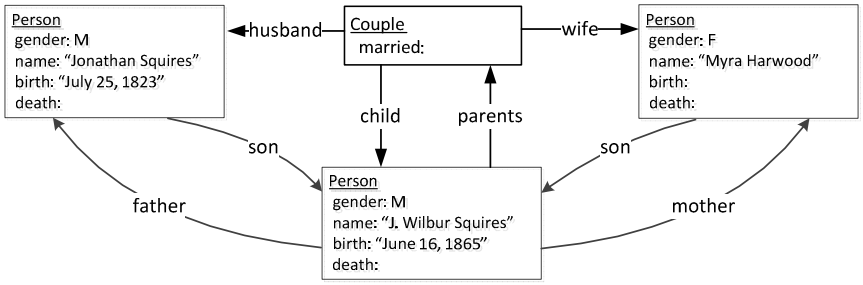




Figure 7: Linguistic Analysis for Case 2

Here the phrases highlighted in the output of the LG Parser are those that require a much deeper linguistic analysis than required for anything in Case 1. Observing the meaning summary graphs in Figures 6 and 7 we see that their structure is exactly the same in spite of the fact that the structure of the original sentences is quite different. This shows the power of the OntoSoar approach.

## Evaluation

One reason for integrating the linguistic analysis of OntoSoar with the OntoES system is that OntoES also includes a number of associated tools. Among these tools is a web-based annotator that allows a person to mark up a page of text with the information on the names, dates, events, and family relationships found there. We intend to measure the accuracy of OntoSoar by comparing its results with human annotation of a randomly selected sample of pages.

## Conclusions

OntoSoar is a powerful tool for analyzing genealogical texts. Through syntactic and semantic analysis and further inferencing we can discover information about people, their life events, and their family relationships. All this information can then be integrated with a conceptual model in OntoES, making a full web of knowledge that will be available to be searched using the web-based tools in OntoES.

## References


Bryant, John E. (2008). *Best-Fit Constructional Analysis*. PhD Dissertation, University of California at Berkeley.

Cimiano, Philipp (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer, New York.

Embley, David W., Steven W. Liddle, and Deryle W. Lonsdale, (2011). "Conceptual Modeling Foundations for a Web of Knowledge", in *Handbook of Conceptual Modeling*, Chapter 15.

Hoffman, Thomas and Graeme Trousdale, eds. (2013). *The Oxford Handbook of Construction Grammar*. Oxford University Press, New York.

Lonsdale, Deryle, Merrill Hutchison, Tim Richards, and William Taysom (2001). An NLP system for extracting and representing knowledge from abbreviated text. In *Proceedings of the Deseret Language and Linguistics Society 2001 Symposium.*

Wintermute, Sam (2012). "Leveraging Cognitive Context for Language Processing in Soar" in *The 32nd Soar Workshop*, June 2012, University of Michigan.

Wong, Wilson, Wei Liu, and Mohammed Bennamoun (2012). "Ontology Learning from Text" in *ACM Computing Surveys*, Vol. 44, No. 4, Article 20.