Towards a Discourse Model for Knowledge Elicitation

Eugeniu Costetchi
CRP Henri Tudor, Luxembourg, 29, J.F. Kennedy, 1855, Luxembourg
Eugeniu.Costetchi@Tudor.lu

Abstract

Knowledge acquisition has been and still remains a hard problem. When it comes to eliciting knowledge from human subjects, an artificial interviewer can be of tremendous benefit. In this paper we present a discourse model for representing the explicit propositional content of a text along with question raising mechanism based on it. This feature is perfectly aligned with the purpose of acquiring more knowledge from the human respondent and acting as a self-extending knowledge base.

1 Introduction

In ontology engineering field, one of the main goals is building an ontology (knowledge base) of a particular domain. The ontology in this case represents a *commonly agreed "specification of a conceptualization"* (Gruber, 1993) within a group of domain experts.

There have been proposed many methodologies to build ontologies e.g. (Ferndndez, Gmez-p, & Juristo, 1997; Noy & Mcguinness, 2000; Uschold & King, 1995). Some are manual, some others are semi-automatic, however, the main burden of interviewing (or eliciting knowledge from) the domain experts, conceptualizing and then encoding the knowledge with a formal language is left on the shoulders of the ontology engineer. Therefore, the process is slow, expensive, non-scalable and biased by the ontology engineer's understanding of the domain.

A solution to knowledge acquisition problem in ontology engineering is envisioned in (Costetchi, Ras, & Latour, 2011). They present a system that could take the role of a human interviewer in the process of knowledge elicitation for the purpose of creating the ontology of the discussed topic. In their vision, one crucial difference to ontology definition is the fact that the created ontology is not shared but it is an *individual* "*specification of conceptualization*" which captures the text propositional content without assuming any prior knowledge of the domain of discourse.

We embark on this idea of artificial interviewer for the purpose of knowledge acquisition as a topic or domain ontology. The proposal is to start from a syntactic and semantic analysis of text (parsing) and interpret the parsed information, through the lens of the Systemic Functional Linguistics (Halliday & Matthiessen, 2004), into a coherent and consistent discourse model. Then it can serve as basis for question generation in order to drive further the knowledge elicitation process. The system, therefore, is intended to act as a self-extending knowledge base by means of written interaction with a human respondent.

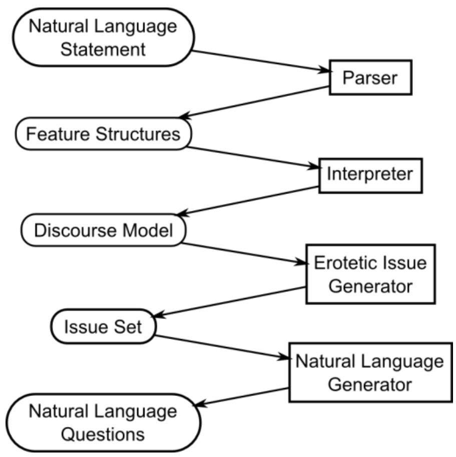

Figure 1: Interaction cycle architecture.

Figure 1 presents the simplified architecture for one interaction. Rounded boxes on the left-hand side represent the data structures; the boxes on the right-hand side represent operational modules and the arrows represent input-output data flows. The *parser* takes natural language text and provides a syntactic and semantic analysis in terms of *feature structures* which are sets of attribute-value pairs. The content of feature structures is systematized according to SFL theory. The *interpreter* instantiates the *discourse model* from the feature structures. The discourse model serves as the central knowledge repository. Based on it and its instantiation the *erotetic issue generator* creates all possible issues that can be raised, given a particular instance of discourse model. An issue is a formal representation of a question. The issues serve as

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 38–44,
Hissar, Bulgaria, 9-11 September 2013.

an *expansion mechanism* of the discourse model. The model extends by accommodating answers (statements) that resolve the issue. Then natural language generator translates formally expressed issues into natural language questions.

The scope of this paper is limited to the discussion of the discourse model and how it can serve as a basis for question raising. Other challenges are just briefly mentioned and left out of the discussion scope.

In next section of the paper is presented the general approach to the problem followed by a section describing the SFL parser. In section 4 we present the discourse model and an example text interpretation. Section 5 provides an example axiomatization employed for question rising which is presented in Section 6. Final remarks and conclusions are drawn in Section 7.

2 The Approach

An interaction cycle between human and system starts with the natural language statement written by human and ends with a set natural language questions generated by the system. The statements are parsed and interpreted in terms of a discourse model which serves as a formal semantic representation of what has been said in the text. The same model serves as a foundation to raise questions (as issues). The raised questions are transformed into natural language text.

For text analysis is employed a systemic functional parser (Costetchi, 2013). It employs a graph-based transformation from dependency parse into a set of feature structures.

The interpretation process consists of instantiating the of discourse model from the feature structures produced by the SFL parser therefore it relied only on linguistic semantics. Pragmatic interpretations like implicatures (Grice, 1975) will not be interpreted as that would require (prior) world knowledge (which is avoided within the system).

SFL adopts a semiotic perspective on language and distinguishes different meaning-lines fused in the text. It provides, among others, linguistic semantics that resembles *frame semantics* (Fillmore, 1985; Minsky, 1974) at the clause level (in terms of processes and their participants) and also *taxis semantics* at the inter-clause level (in terms of logico-semantic relations) which resemble Rhetoric Structure Theory relations (Mann & Thompson, 1988).

To parse in terms of full SFG grammar is computationally unfeasible (Bateman, 2008; Kay, 1985; Robert Kasper, 1988), but it is possible to parse with parts of grammar which provide semantic account of the clause (Costetchi, 2013; Michael O'Donnell, 2012) and inter-clause relations.

The discourse model serves as a foundation for generating questions. If we compare the expansion of the model to a growing plant, then the plant would have buds from which a new leaf, branch or flower can grow. Within the model we define *question raising buds* as "places" in the model where new knowledge can be integrated. And since it is not priory known what that knowledge is going to be, the expansion of the bud is resolved by raising a question and accommodating the answer.

The next section describes the discourse model and provides an example text interpretation.

3 The SFL Parser

The parser (Costetchi, 2013) employs a graph-based approach to generate Systemic Functional Grammar *mood* (chunked functional constituency parse) and *transitivity* (frame semantic account of process type and participant roles) parses from the Stanford Dependency parse (Marneffe, MacCartney, & Manning, 2006; Marneffe & Manning, 2008) and Process Type Database (Neale, 2002). It is a computationally and linguistically viable text parsing approach for natural language which encompasses framed semantic roles together with an adequate syntactic structure to support those semantic roles. An example analysis generated by the parser is presented in Table 1.

example 1	**the duke**	**had**	**given**	**the teapot**	**to my aunt.**
Mood	clause: [mood type: declarative; tense: past perfect simple; voice: active: polarity: positive]				
	subject	predicate		complement	complement
		finite	predicator		
transitivity	agent-carrier	possessive process		possessed	beneficiary
example 2	**the lion**	**caught**		**the tourist**	**yesterday.**
Mood	clause: [mood type: declarative; tense: past perfect simple; voice: active: polarity: positive]				
	subject	predicator/finite		complement	adjunct
transitivity	agent-carrier	possessive process		affected-possessed	temporal location

Table 1: Mood and transitivity example.

The parser produces feature structures representing syntactic and semantic analysis of text. Among the clause *syntactic features* are: *mood, tense, voice* and *polarity* while the clause *semantic features* are the *process type* and *participant roles*. In Figure 2 is presented an example of semantic feature structure.

$$\begin{bmatrix} \text{process type: possesive} \\ \text{process: catch} \\ \text{Ag-Ca: lion} \\ \text{Af-Pos: tourist} \end{bmatrix}$$

Figure 2: Feature structure example.

The parser distinguishes among 16 process types (Figure 3) and 29 participant roles where 17 are simple and 12 are compound. In (Fawcett, 2009) are proposed 65 configurations of process types and participant roles. The semantics of such configurations is captured by GUM ontology (Bateman, Henschel, & Rinaldi, 1995). However the process type and participant role classifications are different, therefore a structural adaptation is required to provide compatibility. We describe the adaptation in the next section.

4 The Discourse Model

The discourse model proposed here draws mainly on GUM. Generalized Upper Model (Bateman et al., 1995) is a linguistically motivated upper level ontology that is domain and task independent. It serves an interface between the linguistic and conceptual forms. This model is compatible with SFL experiential line of meaning which deals with semantic content of text. We further propose a temporal extension and two structural modification of GUM.

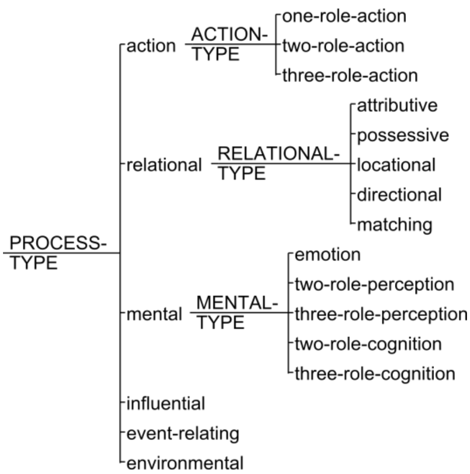

Figure 3: The Process Type classification.

The first structural modification consists in adaptation of process type and participant role classifications. GUM is build based on Hallidayan classification (Halliday & Matthiessen, 2004) whereas we propose to use the one described in (Fawcett, 2009). The main reason for such adaptation is the SFL parser which produces semantic descriptions according to the latter classification. The top level classification of Fawcett's process types is presented in Figure 3.

The second structural modification consists in dividing the process types into *eventive* and *stative* processes. This distinction is metaphysically motivated in DOLCE upper level ontology (Borgo & Masolo, 2009) and linguistically motivated by Bach (1986). This distinction is necessary for the temporal extension of the model. So we propose that *attributive, possessive, locational, emotion and environmental* processes to correspond to states while the *action, directional, matching, perception and cognition* processes to be classified as events. This is an intuitive distinction among the process types based on their description and more fine grained division shall be proposed that will, for example, take into consideration the participant roles as well.

In natural language a finite clause is anchored into the "*here and now*", so to speak, bringing the clause into the context of the speech event. This is achieved either by reference to *the time of speaking* (via tense) or by reference to the *judgment of the speaker* (via modality). We hold the view that, in a narrative, each participant can be described via a temporal evolution complemented by atemporal descriptions (e.g. modal, conditional, causal, concessive, etc.) We focus on the former one and the atemporal one is left for future works.

The temporal dimension provides a linear layout for events and states. Each *participant* has one or more time-lines. The events are distributed along the timeline(s) of the participants. The *events* happen in time and are assumed to be bound by start and end time-points. The *states* last in time and correspond to the conditions and properties of participants along a time interval. They are assumed to be unbound unless a start/end time points and/or duration are specified. Allen (1983) proposes seven basic relations to relate intervals: *before, meets, overlaps, starts, finishes, during and equal.* We incorporate these relations into the model as means to provide a partial ordering to the events and states on the participant timelines. The choice of the temporal relation between two events/states is based on the tense, aspect and temporal circumstances. We do not provide yet a description of the selection conditions but rather focus on motivating their role in a discourse model and in question generation.

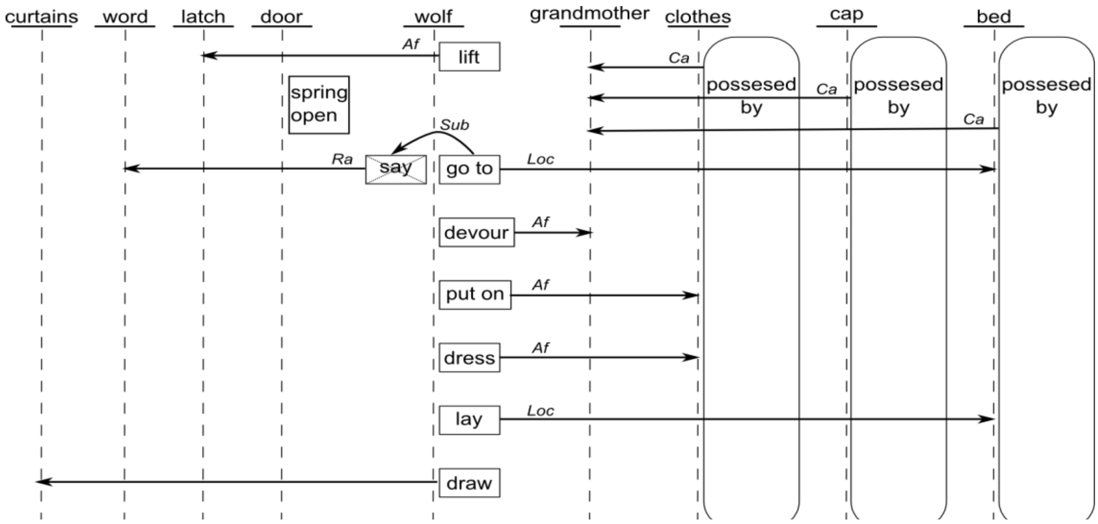

Figure 4: The wolf example from "Little Red Riding Hood".

As mentioned before, not all statements can be integrated into a timeline for they are *atemporal*. For example some present simple clauses cannot be (easily) located in time when they express facts or generalizations. In this case the events are placed on "*atemporal timelines*". In the same manner are treated the conditional or causal relations. The decision to place them on atemporal timelines is merely pragmatic and aims to keep a uniform representation of events and states.

In Figure 4, is provided an example from *"Little Red Riding Hood"*. It is a graphical representation of a paragraph interpreted into the discourse model.

"The wolf lifted the latch, the door sprang open, and without saying a word he went straight to the grandmother's bed, and devoured her. Then he put on her clothes, dressed himself in her cap, laid himself in bed and drew the curtains."

At the top of the schema are all the participants mentioned in the discourse ordered arbitrarily. Each of them has a timeline depicted by a dotted vertical line. The events are drawn by squared boxes while the states by rounded boxes. The events are temporally delimited, positioned and ordered as they flow in the discourse, while the states stretch along the entire duration of the discourse. The temporal interval relations between events are implicit in the graphical representation.

The events are placed on the timeline of the subject participant, e.g. *Agent* that brings about the event or the possessed thing which is the head noun in possessive nominal phrases. Whether it is a state or event (e.g. *wolf lifted the latch*) is decided according to the earlier classification. Note that we treat possessive pronouns in nominal phrases as nominalised possessive processes. For example "*grandmother's bed*" is semantically equivalent to "*grandmother has a bed*" where the *grandmother* is the carrier and the *bed* is the possessed thing.

The participant roles become orthogonal relations from events or states to other participants (and sometimes to events or states, e.g. phenomenon participant role occurring in mental or influential processes). For example in Table 1, the frame semantic relations are the *agent-carrier*, *possessed* and *beneficiary*. So the event of giving is placed on the *lion*'s timeline and from this event there are two orthogonal relations to the *teapot* and *aunt*. Another example is in Figure 4 where *lift* is placed on *wolf's* timeline but it has the second participant *latch* which has the role of affected.

In current model only noun participants are considered. Therefore the pronouns (*he, her*) have to be anaphorically resolved. We assume that there is already a mechanism to resolve anaphora as correference indexing in order to trace the identity of participants and have a concise instance of the model.

This is just a preliminary attempt to characterize the discourse model since it is still a work in progress we do not yet provide a formal characterisation of it.

5 Axiomatization of Process Types and Participant Roles

In SFL, the classification of participants and process types is linguistically motivated. However some common sense principles surface as supporting models. We provide an example axiomatization for a process type and its participant roles. Such axiomatization will also serve as foundation in question generation process.

For example *action* processes are distinguished from *mental* processes as the first one occurs in physical realm while the second one in mental realm. The *actions* are considered to express quanta of change in the world occurring over time and they fall under the *event* category. In other words, the world transitions from an initial state s_i through event e to a final state s_f.

```
Action(e) -> si <before e <before sf
```

The actions can take a limited number of participant roles: *agent, affected, carrier* and *created*. For example *agent* role is given to the participant that brings about the event. We can say that agent x does the action e. The *affected* role is given to the participant that receives some change through action e. The *created* role is given to the participant that did not exist before the action e and it came about as a result of action e. We propose new relations to distinguish between the linguistic semantic and the common sense axiomatization which is of a conceptual nature. Below is the formal expression of relations between participants and the event.

```
Agent(x) -> do(x,e)
Affected(y) -> change(e,y)
Created(z) -> create(e,z)
```

If we put together all the above axioms, we can say that in the world can occur an event which may be a happening (no agent involved) or a doing of an agent. As a consequence there is a state change in the affected participant or creation of a new participant that did not exist before. Also the agent is relevant for pre-event state s_i while the affected and created are relevant for post-event state s_f.

```
Action(e) -> (do(x,e) OR happen(e)) AND
             (change(e,y) OR cre-
              ate(e,z))
```

A similar common sense axiomatization is proposed for *relational* processes. They stand in the opposition to both actions and mental processes and describe the state of affairs. For example, in an *attributive* process, the *carrier* is ascribed an *attribute* which can be either quality, identity, class or an abstract role from the domain model. The attributive processes do not denote any change so they fall into state category. We can say that in a particular state of the world s there is a carrier c that can be characterized by its attribute a.

```
Attributive(s) AND Carrier(c) AND
   Attribute(a)  -> is(c,a,s)
```

Similar reasoning applies to possessive relational processes.

```
Posessive(s) AND Carrier(c) AND
  Posessed(p)  -> have(c,p,s)
```

Now we can say that a state of the world s is characterized by the sum of relations that hold between carriers and their ascribed attributes, possessions, matches etc.

Such axiomatizations fall beyond the discourse model because they are of conceptual nature even if they are derived from a linguistic model. In the next section we describe how questions can be generated from discourse model based on such common sense axiomatizations.

6 On Question Raising

We take a situated and context-bound perspective on knowledge and language. SFL, through semantic frames, provides a linguistic support to situated knowledge while formal representation is provided through situation semantics (Barwise & Perry, 1983).

Given a relation $rel(p_1,p_2 \ldots p_n)$ where all parameters are known we generate an issue by assuming that there exist an alternative value for a parameter p_k where $1 \leq k \leq n$. We formally represent a question via lambda notation as follows:

$$\lambda p_k \; rel(p_1,\; p_2 \ldots p_k \ldots p_n)$$

In the following we illustrate the question rising mechanisms by using as seeds the below examples.

a. *[The wolf]$_{ag}$ [lifted]$_{action}$ [the latch]$_{aff}$*
b. *[grandmother's]$_{car}$ [bed]$_{poss}$*

They can be represented as common sense axiomatization from above, as follows:

```
a. Action(lift) -> do(wolf,lift) AND
                   change(lift, latch)
   si <before lift <before sf

b. Posessive(s) AND
   Carrier(grandmother) AND
   Posessed(bed) ->
             have(grandmother,bed,s)
```

Alternative participant questions are questions aiming to elicit alternative participants given the context of a particular event or state. So we can ask for alternative participants in *do* and *change* relations as follows:

```
λx do(x,lift); λy change(lift,y);
      λc has(c,bed,s)
```

This can be translated into natural language as

"Who else can lift a latch?"
"What else can a wolf lift?"
"Who else has a bed?"

Alternative event questions are questions aiming to elicit in what events the current participants can be in.

`λe do(wolf,e); λe change(e,latch)`

Correspondingly, the natural language expression is:

"What else a wolf can do?"
"What else can happen to a latch?"

State elicitation questions seek to receive new attributes for a given participant:

`λa has(grandmother,a,s)`

"What else does the grandmother have?"

Now taking into consideration change-based axiomatization for actions we can formulate questions about initial and final states even if they are not mentioned in the discourse. To do so we appeal to temporal relations to specify the position of the targeted state relative to the event.

Consequence elicitation questions seek to identify the affected participants and their corresponding post-event attributes. For example if we want to elicit how the latch changed after the event we write it as follows:

$\lambda a \; \texttt{is(latch,a,}s_f\texttt{)} \; \texttt{AND} \; s_f >_{after} \texttt{lift}$

"How is the latch after the lift?" or
"How did the latch change after the lift?"

Temporal elicitation questions aim to elicit new events or states related to a target event. For example, an event e_1 is mentioned in the discourse. Then, for a given an interval relation, e.g. *before*, assume there is an unknown state or event e_2 that stands in this relation to e_1. In natural language, this hypothesis can be translated into a question "*What happened before e_1?*" The satisfiable answer to this question will bring the new event or state statement e_2 into the discourse model. And it will be placed into a *before* relation with e_1.

When decontextualized, the above questions might sound odd or unnatural. Therefore a question selection mechanism would need to be build based on questioning sequences found in natural language dialogues that follow a predictable goal and focus of attention. We do not cover such a mechanism here, but rather are interested to explore means for finding possible question classes. When questions raising methods are clear and the possible classes are known then the selection algorithm can employ them to simulate coherent questioning sequence. So far we have provided some examples of question classes that can be generated from the discourse model, but it is neither an exhaustive nor systematic enumeration of question classes and more work needs to be done in this area. We conclude now on the proposed discourse model and question raising mechanism.

7 Discussion and Conclusions

The current paper is motivated by the idea of an automatic interviewing system. We discuss a preliminary description of a discourse model and question generation mechanism. The discourse model takes as foundation GUM ontology and can represent linguistically motivated semantic relations between entities and events and states in which they participate. However those relations are general enough as to enable further transformation into domain ontology. The model is also temporally imbued so the events and states can be ordered along the timelines of entities. In the last part of the paper we show how questions can be generated for the knowledge elicitation process.

The automatic interviewing system is motivated by ontology building process. The instances of the presented discourse model can be transformed into the topic/domain ontologies once the elicitation process if over. This challenge shall be addressed in the future work.

There are many unaddressed challenges. A few important ones are: reference tracking of participants and events, accommodation of received answers and knowledge update, question selection and sequencing along the interview session, dialogue management and turn taking, natural language generation for questions (either by employing a fully-fledged natural language generation system or a template-based approach suffices for this task).

The discourse model is intended for interactive discourses but it can be employed equally successful on non-interactive discourses with suitable adaptations of the parsing and interpretation modules to the text type. The model could be of prodigious benefit, beyond its intended meaning, for text mining, knowledge acquisition, information extraction, sentiment analysis, expert systems, semantic web and ontology building communities.

References

James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, *26*, 832–843.

Emmon Bach. 1986. The algebra of events. *Linguistics and Philosophy*, *9*, 5–16.

Jon Barwise, & John Perry. 1983. *Situations and Attitudes*. *Semantics A Reader* (p. 352). MIT Press.

John A. Bateman. 2008. Systemic-Functional Linguistics and the Notion of Linguistic Structure: Unanswered Questions, New Possibilities. In Jonathan J. Webster (Ed.), *Meaning in Context: Implementing Intelligent Applications of Language Studies* (pp. 24–58). London, New York: Continuum.

John A. Bateman, Renate Henschel, & Fabio Rinaldi. 1995. *The Generalized Upper Model* . Retrieved from http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/gum/gum-2.pdf

Stefano Borgo, & Claudio Masolo. 2009. Foundational choices in DOLCE. *Handbook on Ontologies*, *2*, 361–381.

Eugeniu Costetchi. 2013. A method to generate simplified Systemic Functional Parses from Dependency Parses. In *Proceedings of DepLing2013 [forthcoming]*. Prague.

Eugeniu Costetchi, Eric Ras, & Thibaud Latour. 2011. Automated Dialogue-Based Ontology Elicitation. *Procedia Computer Science*, *7*, 185–186.

Robin P. Fawcett. 2009. How to Analyze Process and Participant Roles. In *The Functional Semantics Handbook: Analyzing English at the level of meaning*. London: Continuum.

Mariano Ferndndez, Asuncin Gmez-p, & Natalia Juristo. 1997. METHONTOLOGY : From Ontological Art Towards Ontological Engineering. In *Assessment* (Vol. SS-97–06, pp. 33–40). AAAI Press.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, *6*, 222–254.

Paul H. Grice. 1975. Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax And Semantics* (Vol. 3, pp. 41–58). Academic Press.

Thomas R. Gruber. 1993. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. *Formal Ontology in Conceptual Analysis and Knowledge Representation*, *43*, 907–928.

Michael A. K. Halliday, & Christian Matthiessen. 2004. *An introduction to functional grammar*. London: Hodder Education.

Martin Kay. 1985. Parsing In Functional Unification Grammar. In D.Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural Language Parsing*. Cambridge University Press.

William C. Mann, & Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*, 243–281.

Marie-Catherine Marneffe, Bill MacCartney, & Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC 2006* (Vol. 6, pp. 449–454).

Marie-Catherine Marneffe, & Christopher D. Manning. 2008. The Stanford typed dependencies representation. *Coling 2008 Proceedings of the workshop on CrossFramework and CrossDomain Parser Evaluation CrossParser 08*, *1*, 1–8.

Michael O'Donnell. 2012. Transitivity Development in Spanish Learners of English. In *Proceedings of 39th International Systemic Functional Linguistics Conference*. Sydney, Australia.

Marvin Minsky. 1974. A framework for representing knowledge. In P. Winston (Ed.), *The Psychology of Computer Vision* (Vol. 20, pp. 211–277). McGraw-Hill.

Amy C. Neale. 2002. *More Delicate TRANSITIVITY: Extending the PROCESS TYPE for English to include full semantic classifications*. Cardiff.

Natalya F. Noy, & Deborah L. Mcguinness. 2000. Ontology Development 101 : A Guide to Creating Your First Ontology. *Development*, *32*, 1–25.

Robert Kasper. 1988. An Experimental Parser for Systemic Grammars. In *Proceedings of the 12th Int. Conf. on Computational Linguistics*. Budapest.

Mike Uschold, & Martin King. 1995. Towards a Methodology for Building Ontologies. In D. Skuce (Ed.), *Methodology* (Vol. 80, pp. 275–280). Citeseer.