

# On Original Generation of Structure in Legal Documents

Steven O. Kimbrough\*  
Thomas Y. Lee  
Balaji Padmanabhan  
Yinghui Yang  
University of Pennsylvania  
Jon M. Huntsman Hall  
3730 Walnut Street, Suite 500  
Philadelphia, PA 19104-6340  
{kimbrough, thomasyl, balaji,  
yiyang}@wharton.upenn.edu

## ABSTRACT

This position paper advocates a vision in the development of automated legal reasoning and presents evidence supporting the plausibility of that vision. The paper observes that original creation of documents of legal import in either fully formal or semistructured form offers the prospect of greatly reducing the cost and expanding the scope of knowledge engineering for legal reasoning. This, it is claimed, is most likely to be achieved via formalization of various sublanguages of legal discourse. SeaSpeak is an example of such a sublanguage and it appears to be amenable to full formalization. Short of that, much can be done with partial formalization and semistructured documents. The paper presents a tabular format for message expression, motivated by a formal agent communication language.

## 1. INTRODUCTION

Automated reasoning, however intelligent, needs something to reason upon, a formalized knowledge base of some kind. AI in support of legal reasoning is no exception. Here, it has been an enduring challenge to find ways of obtaining sufficiently structured documents. In other domains people may be the primary targets of knowledge engineering; in AI and the law much of the requisite knowledge resides in documents of various sorts.<sup>1</sup>

Compromising severely in favor of brevity over accuracy, there have been three main approaches to extracting formalized knowledge from documents of legal interest.

1. *Manually symbolize a relevant corpus.* The approach

\*Corresponding author.

<sup>1</sup>See [31, 43] for very useful reviews of knowledge-based systems for legal reasoning.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

here is to pick an appropriate formal representation language and manually (perhaps with computerized support) symbolize the essential information from the relevant collection of documents, typically statutes, administrative rules, and legal cases. Sergot et al. [44] and Bench-Capon et al. [2] are early examples of this approach; [7, 20] are recent examples.

2. *Accept minimally structured documents.* Under this approach, the documents in the chosen corpus are formalized only in a very weak sense, e.g., by creating inverted files of their terms and limiting inference to what Information Retrieval techniques can produce. Choueka et al. [5] is an early example; the practice is by now, of course, ubiquitous. Closely related is conceptual information retrieval, e.g., [9], as well as hybrid approaches, e.g., [38], that rely upon a modest amount of (usually) manual structuring. Information extraction techniques [4] have also been applied successfully in the legal domain [14, 27].
3. *Automatically structure a relevant corpus.* Under this approach, pattern-finding programs extract structure from documents in a target corpus and create derived documents, often in XML, which present their structures more transparently. Interesting examples include [30, 40, 41].

These three approaches (and their combinations) have complementary strengths and weaknesses. Manual symbolization affords the best prospect for deep and detailed automated inferencing and information recovery, yet it is the most labor intensive option and presents serious problems of maintenance. Accepting minimally structured documents is the least expensive alternative and the least powerful in terms of potential to support inferencing. Automated structuring lies more or less between the other two approaches. The three may be thought of as defining an operating curve that trades off cost and inferential acuity, much as gearing on a bicycle presents an operating curve that trades off speed and power.

Are there any ways to change the location of the operating curve itself, rather than limiting ourselves to seeking the best point on it for a given application? Indeed there are such ways. Our aim in this paper is to discuss part of one such

family: arranging so that documents *as originally created* have the requisite structure to support automated inferencing for a given application. The thought is to create these structured documents as *by-products* of normal operations.<sup>2</sup> The documents may be semistructured or fully-structured. The former is a popular idea and we shall discuss it only briefly. The latter is relatively under-addressed and we shall discuss it in more detail. The idea we wish to explore is that, in restricted domains of import to systems of law, it may be feasible to impose special-purpose (artificial sub-) languages and lexicons, and that doing so will materially be helpful in addressing the challenge of obtaining sufficiently structured documents for purposes of automated reasoning. We turn first, then, to short discussion of artificial sublanguages.

## 2. ARTIFICIAL SUBLANGUAGES

Language enables communication. Languages inhibit communication, for communication requires a common language and the cost of learning multiple languages raises an often unsurmounted barrier.<sup>3</sup> Having a *lingua franca*, a general language known (more or less) universally, would afford more or less universal communication. At various times and places certain natural languages, such as Greek, Latin, Mandarin, and French, have served broadly, if not universally, as common communication vehicles.

Today English in some form appears headed towards being the universal language of commerce and affairs. The fact remains, however, that universal proficiency in English is not around the corner. Further, even with universal fluency in English there are, and will always be, realms of discourse for which precise and accurate communication is required concerning specialized topics. It is not enough to have basic knowledge of English if the purpose of communication is air traffic control, navigation, law enforcement, and so on. In these and many other realms of discourse there exist specialized concepts and vocabulary that have to be mastered in the interests of efficient and effective communication.

Specially-designed languages can in principle be created that are relatively easy to learn and that are sufficiently expressive for particular purposes. They need be mastered only by a given community of interest. This idea has had an extensive history and considerable uptake, and it goes by a number of names. Besides *planned languages* the literature also uses the terms

artificial languages, constructed languages (conlangs), invented languages, imaginary languages, fictional languages, etc., including universal languages, auxiliary languages, interlanguages or interlinguas, international languages; and also including logical languages, number languages, symbolic languages, etc. [12]

as well as others.<sup>4</sup> Our focus is on artificial languages that

<sup>2</sup>The idea is hardly new or originated by us. It has in fact appeared at ICAIL in previous years, e.g., [6].

<sup>3</sup>We note the possibility that the net effect of multiple languages may be to foster communication, perhaps because there is little advantage in communicating if everyone can understand you, or because specialized languages make communication more profitable in niche applications. But these are matters beyond the scope of the present paper.

<sup>4</sup>There is even a hobbyist Web site. See <http://www.langmaker.com/ml0101.htm> for a manifesto!

are also *restricted languages* or (the term we shall use) *sublanguages*.

Informally, we can define a sublanguage as the language used by a particular community of speakers, say, those concerned with a particular subject matter or those engaged in a specialized occupation.[39, page 2]

Several artificial sublanguages have been fielded, are successfully in use today, and are not without relevance to legal reasoning, broadly construed.<sup>5</sup> Examples include AirSpeak, SeaSpeak, PoliceSpeak, and LinguaNet (see [3, 15, 16, 35, 28, 17] for an overview). These languages were designed to be easily learned so that they can be spoken and heard effectively. Their employment and continued development today suggests they will be useful in the longer term. They employ a controlled, or restricted vocabulary. As is typically true with sublanguages, their

... grammar contains additional rules not satisfied by the language as a whole. It also happens that some of the grammatical rules of the language as a whole disappear, i.e., do not apply, in a sublanguage. Since the sublanguage must satisfy the rules for the language, this disappearance is possible only if the rules are satisfied vacuously in the sublanguage, i.e., if certain word classes or well-formed sequences or transformations do not appear in the sublanguage. [11, page 154]

*Telegraphic languages*, yet more austere forms of sublanguage, are also widely in use and readily display the simplified, constrained grammar typical of sublanguages. We note that *telegraphic language* and *telegraphic speech* are also terms of art in the field of child development, and it is here that the terms obtained their original meaning.

When children initially produce grammar, their language often sounds rather like the abbreviated language of telegrams (“Daddy gone,” “Mummy shoe.” “See big car”). This is why, in the past, this type of early output was referred to as telegraphic speech. At this stage, toddlers omit indefinite and definite articles, as well as prepositions and the like. They also leave out morphemes like plural “s,” progressive “ing,” and possessive “s.” [19, page 94]

Telegraphic languages are not unknown among adults. Fitzpatrick et al. [10] present a particularly clear case study of a telegraphic language used in the U.S. Navy. The stylization apparent in examples from this language—e.g., “72 man-hours expended,” “Stock requisition shipped,” “Work request submitted,” “Improper repair work performed,” “No parts required” [10, page 45]—will be familiar to the reader.

Can working artificial languages—especially sublanguages and telegraphic languages—be formalized so that machines may productively conduct inferences using them? This question is interesting both theoretically and practically. Theoretically, the question presents an apt challenge for ACLs

<sup>5</sup>We shall not discuss various more ambitious efforts to develop general-purpose universal languages. Esperanto is perhaps the most well-known candidate language. For relevant background see [25, 29, 32, 37, 45]; also, Harrison [12] has put together a very useful bibliography.

(agent communication languages), including the various projects to create FLBCs (formal languages for business communication), and the various XML representation efforts. Can the ACLs adequately represent a given artificial language? If not, how might they be improved? What does formalization of artificial languages tell us about requirements for ACLs? From a practical point of view, formalization could afford human-machine and machine-human communication, including language translation and error detection, as well as machine-machine communication, with its attendant possibilities of reducing time and labor costs. Perhaps of most immediate use, formalization and structuring present opportunities for automated recovery and discovery of information.

### 3. ON FORMALIZING SEASPEAK

SeaSpeak is a carefully articulated, if informal, sublanguage for maritime communication. We are engaged in constructing a formal, if approximate, representation of it using Kimbrough's FLBC, a particular agent communication language (ACL). We report here indicative results. Our main purpose is to discuss SeaSpeak as an example of a sublanguage amenable to formalization.

#### 3.1 Background on SeaSpeak

SeaSpeak is also known as "English for maritime communications." It is the language of merchant shipping, a restricted, English-like language adopted in 1988 by the International Maritime Organization (IMO) of the United Nations for use in ship-to-ship and ship-to-shore communications. Greatly increased shipping during the 1960s and 1970s and the fact, that the distribution of nationalities of ships' officers and crew changed from roughly 80% English-speaking and 20% other to roughly 80% other and 20% English-speaking, motivated its adoption. The need for regularization of practices in one language and the training of officers in its use was therefore agreed to, and English, already the language of civil aviation, was chosen by the IMO.

SeaSpeak was created by specialists in maritime communications and applied linguistics [16] during 1982–3. SeaSpeak is a system for speech communication, and it is intended for use in situations where it is essential that communication should be as clear, brief and accurate as possible.

SeaSpeak regulates ways of speaking, ways of establishing a conversation, and defines technical vocabulary. All messages begin with a message marker that indicates the nature of what follows, such as advice, information, instruction, intention, question, request, warning, or a response to one of these. The definitive reference for SeaSpeak is the *SeaSpeak Training Manual* [46], upon which we draw for our analysis and formalization.

From a formalization perspective the central concept in SeaSpeak is that of a message marker. The manual has this (and not much else) to say about message markers in general [46, page 96].

Maritime messages transmitted over VHF should be short, accurate, and relevant. Furthermore, messages should be transmitted in language simple enough for a non-native speaker of English to comprehend without difficulty.

One useful means of making the language simpler is to indicate, at the beginning of a message,

what sort of message it is going to be. Thus, if a question is going to be asked, the speaker simply says the word 'QUESTION' before the question itself. Similarly, if a piece of advice is going to be given, the speaker says the word 'ADVICE' in advance of his message. There are just seven of these *Message Markers* and after a little practice, learners should experience no difficulty in using them.

These *Message Markers* have another function: that of imposing order on the conversation, since each message marked in this way requires a reply correspondingly marked (even if that reply is nothing more than an acknowledgement of the message received). This procedure helps to ensure that:

1. messages do not become confused with each other
2. each message is dealt with in turn
3. a participant receiving a reply knows which message is being replied to.

SeaSpeak has only seven markers (with a mirroring reply-marker in each case). The seven are [46, pages 96–7]:

1. Information (Information-Received)
2. Warning (Warning-Received)
3. Intention (Intention-Received)
4. Request (Request-Received)
5. Advice (Advice-Received)
6. Instruction (Instruction-Received)
7. Question (Answer)

SeaSpeak's message markers are, as we shall analyze them, essentially speech act operators or illocutionary force indicators. The content that they govern is simple in form and not entirely specified and closed. Here is a summary from the training manual [46, page 160].

1. SeaSpeak messages are formed entirely from words within the English language.
2. The total vocabulary used in SeaSpeak comprises 3 kinds of words and expressions:
  - (a) **The vocabulary of 'general' English.** Knowledge of the non-specialized vocabulary of English is assumed, and so it is not listed in the SeaSpeak Vocabulary.
  - (b) **Words in general maritime use.** These words occur frequently in maritime communications, and are listed in Section I, as Categorized General Maritime Vocabulary.
  - (c) **Words in specialised maritime use.** These words and expressions may occur only rarely in general maritime use, but frequently in particular circumstances or for specific communication subjects. They are listed in Section II under the Major Communications Subjects.

Item (2a) presents a particular challenge to formalization efforts. Just what is the scope of 'general English'? Examples are useful. Here and in the sequel we will use *italic font* for messages in SeaSpeak. The following examples are from [46, page 97].

1. *QUESTION: What is your ETA at the dock entrance?*
2. *INSTRUCTION: Go to berth number: two-five.*
3. *ADVICE: Anchor, position: bearing: one-nine-four degrees true, from Keel Point distance: one mile.*
4. *REQUEST: Please send, quantity: five acetylene cylinders.*
5. *INFORMATION: The pilot is waiting now, position: near buoy number: two-six.*
6. *WARNING: Buoy number: two-five and buoy number: two-six are unlit.*
7. *INTENTION: I intend to reduce speed, new speed: six knots.*

SeaSpeak’s carefully-designed and circumscribed structure favors formalization. Two general issues, however, present substantial challenges, both in SeaSpeak and elsewhere: illocutionary force indicators and fixing of reference. SeaSpeak affords a good context for discussion of these topics.

### 3.2 Illocutionary Forces

SeaSpeak’s message markers are examples of *illocutionary force* (or *speech act*) indicators [23]. Typically, and perhaps nearly always, a meaningful utterance may be analyzed as an illocutionary force applied to a propositional content. This is the so-called  $F(C)$  framework, due to Searle and Vanderveken [42]. SeaSpeak fits the framework easily and directly. The larger issue is how to formalize illocutionary force indicators. This has been a main topic throughout the development of ACLs and of Kimbrough’s FLBC in particular (see [21, 22, 23, 24]). SeaSpeak presents an apt test for the FLBC approach to representation of illocutionary forces and speech acts. The problem is that illocutionary forces are generally thought to create intensional contexts. For example, to assert, or to promise, or to request, etc. that  $P$  is not to assert, or to promise, or to request, etc. that  $Q$ , even if  $P$  and  $Q$  are equivalent. This is a difficult topic and one beyond the scope of the present paper. We assume a treatment along the lines described in [22], but nothing much turns on the assumption.

### 3.3 Reference Fixing

Referring expressions in ordinary language are usually not proper nouns. For example, in *The cat is on the mat* reference is made to a particular cat and a particular mat. Presumably the cat has a (proper) name, and the mat does not. More usually, instead of proper names, in ordinary conversation we use *indexical expressions* to indicate or fix reference. These include pronouns (*he, she, it, they, etc.*), definite and indefinite descriptions (*the cat, a cat, etc.*), demonstratives (*that cat, those cats, etc.*), and nonverbal pointing (e.g., nodding in a certain direction, waving a hand, pointing a finger). SeaSpeak messages, as evidenced by the examples above, make ample use of reference fixing devices not based on proper nouns. This presents a challenge for any formalization effort.

### 3.4 Structure of SeaSpeak

SeaSpeak evidences a fairly simple, hierarchical governing structure. At the most general, encompassing level there is an *exchange procedure* that governs all conversations. Typical steps include: make contact, agree and switch to a suitable working channel, exchange messages, and terminate. There are detailed rules, e.g., for what happens when contact is abnormally ended during an exchange and for what are proper

responses to which types of messages. Under the exchange procedure are rules for *message specification*. These are, for present purposes, the main focus of our attention.

Speech act theory finds an  $F(C)$  structure underlying all utterances, where  $F$  is an illocutionary force and  $C$  is its propositional content. SeaSpeak, as we have seen in the examples above, broadly conforms to speech act theory in this regard. The 7 (+7) message markers indicate illocutionary force and may roughly be interpreted as follows:  $F(C) \rightsquigarrow$  *Speaker s Fs addressee a that C*. Thus, we have a subordinating construction: *F that C*. Specifically, but approximately,

1. Information: *s informs a that...*
2. Warning: *s warns a that...*
3. Intention: *s announces to a x’s intention that...*
4. Request: *s requests of a that...*
5. Advice: *s advises a that...*
6. Instruction: *s instructs a that...*
7. Question: *s questions a regarding...*

SeaSpeak message contents—corresponding to  $C$  in  $F(C)$ —are simple, consisting of one or two sentences (at most), semantically constrained and often quite stylized. Message markers are not iterated, so we do not find constructions such as *REQUEST(INSTRUCTION(C))*. Content sentences may, however, be at times complex, having—or naturally interpreted as having—a subordinating construction. For example, in the exchange

*INSTRUCTION: Stop immediately.*  
*INSTRUCTION-RECEIVED: Stop immediately,*  
*negative: reason: I am towing now.*

The respondent is saying, roughly “I will not stop immediately, because [subordinator] I am towing now” or “I will not stop immediately. My reason that [subordinator] I will not stop immediately is that [subordinator] I am towing now.”

## 4. ELEMENTS OF FLBC

FLBC in our context (see, e.g., [21, 23, 24, 22]) is an open and evolving research program aimed at developing formal languages for business communication that:

1. are expressed (insofar as possible) in first-order logic<sup>6</sup> (call this the *FOL* aspect)
2. assume a variety of *event semantics* for the target languages being represented (here, SeaSpeak; call this the *ES $\Theta$*  aspect), and
3. recognize and represent speech acts (call this the *speech act* aspect), and other phenomena that create intensional contexts,

for the analysis of languages for business communication. In the remainder of this section we discuss individually the *ES $\Theta$*  and the speech act aspects of FLBC, presenting a brief tutorial and providing serviceable principles of representation.

<sup>6</sup>That is, the default methodological assumption is that first-order logic will be used and that a certain burden of proof must be overcome in order to employ extensions to first-order logic.

## 4.1 ES $\Theta$ : Event Semantics + Thematic Roles

ES $\Theta$  theory has been developed by a number of authors as a way of doing semantics for natural language (see for starters [8, 13, 26, 34]). There is no unified, generally accepted theory. Instead, there are various partial theories and investigations, united by a shared intuition or perspective. At the level of a slogan we would describe the shared intuition as *words categorize things, and are largely about underlying events, states, and processes*. The upshot of this idea is best seen through examples, which we will present shortly.

Events, states, and processes are conceptual primitives in the ES $\Theta$  world view. As concepts they merit and have received much attention, which we here mostly eschew. Intuitively and approximately:

- Events are happenings and normally occupy, or *culminate* at, a moment in time.

Examples: an arrival, a departure, an approval.

- States are beings, properties or conditions that normally *hold* for an extent of time.

Examples: being a student, waiting.

- Processes are happenings that extend in time; after they *commence* they *hold* for some time; they start and stop, but may or may not finish, or *culminate*.

Examples: getting a Ph.D., going to New York.

In the absence of satisfactory definitions of these concepts there can be, and in fact is, often good agreement on how they are distinguished in practice, especially the practice of using language. A terminological note: by *eventuality* we mean something that is either an event, a state, or a process. (They are assumed to be disjoint; the term is Parsons's [34].) Because it is a bit of an ugly word, we shall at times use *event* for *eventuality*, leaning on context for disambiguation.

## 4.2 Simple Sentences

Simple sentences do not involve speech acts or other features that create intensional contexts. Our examples start with and then build on simple sentences. Below, we organize the examples by parts of speech in ordinary English: verbs, adverbs, nouns, and adjectives. Verbs are typically about—make reference to—underlying events, according to ES $\Theta$  theory. Some verbs, however, are about states and still others are about processes.

### 4.2.1 Verbs of event

Most verbs are interpreted as referring to underlying events. Such verbs as *arrive*, *depart*, *reduce*, *stab* can normally be understood as implicitly referring to underlying events. *There is an arrival* has as a stylistic variant (in ES $\Theta$  theory) *There is an arrival event*. Formally, we have  $\exists e(\text{arrive}(e))$ . It shall be our practice to eliminate existentially quantified variables in favor of individual names. Formally, this gives us *arrive*( $e_1$ ), assuming that  $e_1$  is a name constant for an event, and that the speaker has an unlimited supply of (subscripted) unique event names available.

Less minimally, we interpret *John arrived* as equivalent to (as stylistic variant of) *There is an arrival event, it happened before the present moment, and its theme (what it was that arrived) is John*. Formally then  $\text{arrive}(e_1) \wedge \text{Theme}(e_1,$

*John*)  $\wedge \text{Cul}(e_1, t_1) \wedge t_1 < \text{now}$ . *Cul* is a predicate for indicating when an event culminates or happens. *Theme* is an example of a *thematic role*. These are generic predicates used for modifying eventualities associated with verbs. For an intransitive verb such as *arrive*, *Theme* is used to indicate the subject of the sentence, what it is that arrived in this case. When the verb is transitive, *Theme* is typically used for the direct object. Thus *a gave b to c* would be symbolized as  $\text{give}(e_1) \wedge \text{Agent}(e_1, a) \wedge \text{Theme}(e_1, b) \wedge \text{Benefactive}(e_1, c)$ . *Agent* is a thematic role predicate for the (active) subject for a transitive verb and *Benefactive* is a thematic role predicate for an indirect object that is a person (in an extended sense, including corporations). Thematic roles afford much economy of expression. We shall use them without dwelling on the fine points of the associated theory. See the ES $\Theta$  literature for that.

### 4.2.2 Verbs of state

Some verbs are most naturally understood as making reference to underlying states. Examples include *to hunger* (*Jill hungers* is interpreted as *There is a hungry state and Jill is in it*), *to be*, *to have*, *to love*, *to wait* and *to understand*.

*John waited* posits a state of waiting and says that John was in it. In FLBC:  $\text{wait}(s_1) \wedge \text{In}(\text{John}, s_1) \wedge \text{Hold}(s_1, [t_1, t_2]) \wedge t_2 < \text{now}$ . The predicate *In* may be taken as a thematic role, applicable to subjects of verbs of state. *Hold* should be interpreted as saying that a state holds or obtains during the  $[t_1$  to  $t_2]$  interval. (Compare with [18, page 503].)

### 4.2.3 Verbs of process

Example: *John goes home* posits a process undertaken by John. He may go home, yet never arrive. Again following the lead of Parsons [34], we interpret processes as being about underlying events that *hold* over a given period of time. Note: some verbs may be of events in one form and of processes in another: *arrive* (events) and *arriving* (processes). *The plane arrived* implies that it got to the destination in question. *The plane was arriving* does not imply (sadly) that it actually got in.

### 4.2.4 Adverbs

Example [34]: *Brutus stabbed Caesar violently* is implicitly saying that *There was a stabbing event, that event was perpetrated by Brutus on Caesar, and the event was violent*. Symbolically:  $\exists e, t(\text{stab}(e) \wedge \text{Agent}(e, \text{Brutus}) \wedge \text{Theme}(e, \text{Caesar}) \wedge \text{violent}(e) \wedge \text{Cul}(e, t) \wedge t < \text{now})$ .

Prepositions and other qualifiers are treated similarly. Example: *The Andrea Doria is at Keel Point* is implicitly saying that there is a state of being, *s*, which is at Keel Point, and the Andrea Doria is in it. Formally, in FLBC,  $\exists s(\text{being}(s) \wedge \text{In}(\text{Andrea Doria}, s) \wedge \text{at}(s, \text{Keel Point}))$ .

### 4.2.5 Common nouns

Example: *XBar-Harbor-B is an anchorage* is implicitly saying that there is a state of being an anchorage and XBar-Harbor-B is in it. Formally, in FLBC,  $\exists s(\text{anchorage}(s) \wedge \text{In}(\text{XBar-Harbor-B}, s))$ .

### 4.2.6 Adjectives

Example: *The Andrea Doria is late* is implicitly saying that there is a state of being late and the Andrea Doria is in it. Formally, in FLBC,  $\exists s(\text{late}(s) \wedge \text{In}(\text{Andrea Doria}, s))$ .

### 4.3 An Example: SeaSpeak’s INFORMATION

Consider the simple SeaSpeak sentence:

SEASPEAK SENTENCE 1. *INFORMATION: No vessels are at the anchorage.*

We begin by analyzing the simple sentence, *No vessels are at the anchorage*, deferring briefly discussion of the illocutionary force indicator, *INFORMATION*. Thus,

SEASPEAK SENTENCE 2. *No vessels are at the anchorage.*

Note that *the anchorage* is a referring expression whose meaning cannot be recovered from the bare sentence. We assume for the sake of the example that a definite anchorage has been identified, having proper name *XBar-Harbor-B*.

The first step in formalizing a given SeaSpeak sentence is to convert it to more transparent forms, while remaining in natural (informal) language. The alternative forms are called *stylistic variants* and are intended to be adequately similar in meaning (for the purposes at hand) to the original sentence. We begin by eliminating referring expressions in favor of the appropriate proper names. The first stylistic variant is thus:

SEASPEAK SENTENCE 3. *XBar-Harbor-B is an anchorage and no vessels are located at XBar-Harbor-B.*

Notice that sentence 3 is a clearer version of sentence 2. Our next stylistic variant is rather stilted. Even so, it is a natural result from the perspective of event semantics. We need one further transformation:

SEASPEAK SENTENCE 4. (1) *There is a state of being an anchorage,  $s_0$ , and XBar-Harbor-B is in it.* (2) *There is a state of being,  $s_1$ , whose location is XBar-Harbor-B.* (3) *There is a state of being a vessel,  $s_2$ .* (4) *Nothing now in the vessel state,  $s_2$ , is in the state  $s_1$ .*

The function of clauses (1), (2) and (3) is to fix the references of the terms under discussion. (4) belongs to the content part and needs to be wrapped with the message marker. We separate these two functions and distinguish the reference-fixing part of the message from the message body proper.

#### 1. Reference-fixing

##### (a) Declarative

$s$  = the speaker

$a$  = the addressee

$a_1$  = *XBar-Harbor-B* (the name of a particular anchorage)

##### (b) Substantive

$anchorage(s_0) \wedge In(a_1, s_0)$

*XBar-Harbor-B is an anchorage.*

#### 2. Presupposition

[Not used in this example.]

#### 3. Message body

$\forall x \forall s (vessel(s) \wedge In(x, s) \wedge Hold(s, now))$

$\rightarrow \neg Located(x, a_1)$

Call this  $\phi$ . Further, the utterance has the message marker *INFORMATION*. So,

$$information(e_1) \wedge Speaker(e_1, s) \wedge Addressee(e_1, a) \wedge$$

$$Cul(e_1, now) \wedge Object(e_1, [\phi])$$

We treat *information* as a form of assertion, which may be veridical or not. We will have a general rule to the effect that information events are veridical if and only if their objects, here  $\phi$ , are true. See [22] for an elaboration.

We have had success in formalizing example messages of all seven message marker types. We believe, based on our experience working with SeaSpeak, that a substantial and reasonably comprehensive subset of SeaSpeak can be straightforwardly symbolized using the FLBC apparatus sketched above. This remains to be demonstrated, however. A particular advantage of  $ES\Theta$  theory in the present context is its reuse of predicates—especially the thematic roles and common prepositions—and their direct mappability to tabular form (see §7). Finally, SeaSpeak is a good test case. It is a live sublanguage, used in critical conditions. If it can successfully be formalized, or even substantially formalized into semistructured documents, surely other sublanguages of legal interest will also succumb to analysis.

## 5. SEMISTRUCTURED DOCUMENTS

As noted in §1, from the perspective of information recovery, legal documents reside along a continuum from unstructured free-text to highly-structured information represented in a formal, symbolic language. In between these two poles lies the broad range of semistructured documents, for which formal theories of management and manipulation are still evolving [1]. In this section, we consider the relationship between sublanguages and the generation of semistructured documents. We begin with a brief review of the semistructured data model. We then revisit our discussion of sublanguages, this time from the perspective of the semistructured data model. Examples of familiar semistructured data models are cast as sublanguages to illustrate the generation of semistructured documents.

### 5.1 Semistructured data

Semistructured documents are collections of data that conform to some instance of a semistructured data model in the same way that the tuples of a relational table represent the extension of some intensional relation defined on the relational data model. The semistructured data model is most simply defined as self-describing, or slightly more formally, as collections of label-value pairs [1]. In the simplest formulation, if  $l$  is a node and  $v$  is a value (as in a label-value pair), then we can model semistructured data as a tree— $l:v$  |  $l:tree$ —where *tree* denotes an unordered list of sub-trees.

Taken in this context, an HTML document is an instance of a semistructured data model that captures information about the structural elements from which documents are formed, certain relationships between those structural elements, and certain characteristics about how those elements are visually presented. For example, we know the title of the document is the value associated with the HTML label (tag) “title”. We know that the value (text) associated with an “H2” label bears a part-of relationship to the text of the corresponding “H1” label in which the H2 content is nested. Furthermore, we know that text associated with a “Center” label is to appear as centered within some presentation field.

## 5.2 Sublanguages and semistructured data

Revisiting §2, we can think of a sublanguage as a specialized combination of a lexicon and a grammar. A lexicon identifies a vocabulary and its corresponding definitions. The grammar defines how elements of the lexicon may be composed to form valid statements in the sublanguage. In the context of semistructured data, the lexicon defines the set of possible labels and their corresponding semantics. The grammar constrains how labels relate to one another.

As noted in §3, not all sublanguages are complete, viz. “The content that they govern is simple in form and not entirely specified and closed.” The vocabulary might be incomplete. There might be concepts or content that cannot be represented in the lexicon and/or grammar.

From this perspective, we can see how HTML constitutes an incomplete sublanguage. The tag set is both predefined and has a fixed semantics. The rules of well-formedness constrain tag nesting. Much (arguably all) of the information in an HTML document is contained not in the labels but in the free text of the terminal leaf nodes.

While the eXtensible Markup Language (XML) is not itself an instance of a semistructured data model, it is a metalanguage for defining semistructured models. The Document Type Definition (DTD) for an XML instance such as ebXML, however, is one example of an incomplete language for describing the terms used in electronic business and their corresponding relationships to one another. Because even ebXML is incomplete, leaf nodes in these semistructured models allow for free text to capture concepts and relationships not encoded in the formal vocabulary and grammar.

A complete language in the semistructured context, then, is one in which all information is captured in labels and all terminal values are empty. A complete language would eliminate the need for free text. As an aside, we note that any incomplete sublanguage could be interpreted as a complete language by simply discarding free text in label values of the corresponding semistructured model.

Revisiting the example of §3, SeaSpeak defines seven labels (and their corresponding response markers) that can appear as the root of a semistructured document. Other terms from both specialized and general maritime use are defined. Finally, SeaSpeak is an example of an incomplete language because it does ultimately allow for the inclusion of terms from the general, English vocabulary.

## 6. EXPLOITING STRUCTURES

Once the structure of legal documents is made explicit, be it semistructured or fully formal, there are interesting data mining possibilities. For example, consider a SeaSpeak conversation as an example of a specific type of legal document and imagine we have a substantial corpus of such documents. As we have seen given the nature of this document, any reasonably detailed formal or semiformal structuring of such a corpus is likely to be complex and hence the number of structured elements can be large. Some of the common elements of this structure would include names of parties, actions taken, actions requested, and information about outcomes. There are at least three types of data mining opportunities:

1. *Building specific data mining models.* This can help develop strategies for handling new situations that arise, for finding profiles of problematic situations, for detecting suc-

cessful and unsuccessful policies, and so on. For example, given a new case or situation, elements from the case as it unfolds can be used to retrieve similar cases. Among these cases, various classification models, such as C4.5 (Quinlan [36]) can be built to distinguish conditions that result in a favorable outcome with those that result in unfavorable outcomes. For example, the classification model built may show that if there is a specific set of actions taken, then favorable outcomes usually result. This is clearly operationalizable.

2. *Learning interesting patterns.* It would be interesting if we find that even when the recommended actions are taken, if the case is a medical emergency in the North Atlantic, then bad outcomes tend to result. This is an example of a pattern in data, and there has been much work done in techniques that can learn “interesting” patterns from data (e.g. Padmanabhan and Tuzhilin [33]). These patterns may be pieces of potentially useful information, and are different from classification models in that they are more general. Other examples of interesting patterns may be those that suggest novel lines of legal reasoning.

3. *Querying for explicit items.* Here is a partial list of potential queries that would be possible were fully formalized or semistructured corpora available in scale and in scope.

1. We could trace the evolution of a legal concept by calculating the transitive closure of references to the concept (in other words, the grammar should include explicit formalisms for the nature of references - like specialization, generalization, combination, subdivision - see UML) For example, applying the 1921 Martin Act, a broad, New York State law barring fraud in the sale or offering of securities. Eliot Spitzer, New York Atty General is using this to prosecute and regulate Wall Street.
2. We could check the quality and consistency of documents by defining the concepts of well-formedness and validity with respect to the lexicon and grammar.
3. We could check the quality and consistency by defining additional constraints (inclusion, inverse, domain, etc.).
4. We could support robust, flexible querying through a corpus of decisions and legal documents.
5. We could exploit active database concepts (event-condition-action) rules to automatically update indexes to support robust querying.
6. We could automatically update references among legal decisions and policy documents.
7. Given a well-defined relationship between concepts (see below), we could search explicitly for novel instances of a concept within a legal domain for novel lines of legal reasoning.
8. We could search for relationships within dissimilar legal domains for parallel relationships to uncover novel lines of legal reasoning.

## 7. SEASPEAK REPRESENTATIONS

Specifically, how might these ideas be put to use? We present in this section a series of examples of SeaSpeak messages formalized in a tabular fashion. We believe that the transformations to and from FLBC (and in general, to and from an adequately expressive and accurate ACL) are more or less straightforward. If so, then the tabular representation has rather obvious promise as a representation for end

users. Further, the tabular form affords partial structuring while retaining the capacity to be mapped to and from a semistructured format such as XML. Thus, short of full formalization, the ideas in FLBC (event semantics, thematic roles, etc.) remain useful and can be exploited by non-logical techniques, as discussed above.

## 7.1 INSTRUCTION

We begin with the simplest of examples: *Paisano, this is Shell Southport. INSTRUCTION: Go to berth number: two-five. go* is here used in WordNet’s first sense for the verb:

1. travel, go, move, locomote – (change location; move, travel, or proceed; “How fast does your new car go?”; “We travelled from Rome to Naples by bus”; “The policemen went from door to door looking for the suspect”; “The soldiers moved towards the city in an attempt to take it before night fell”)

The message may be represented in tabular format as follows:

To: Paisano	From: Shell Southport
INSTRUCTION	
go	
Experiencer	Paisano
Source	–
Goal	(berth, number: two-five)
Time	–

*INSTRUCTION(C)*, here and in SeaSpeak generally can be interpreted as *Speaker invokes its authority to request that addressee see to it that C*. What is being asked for is that there is a going of Paisano to berth 25. When and from where is left unspecified, but could be stated in this tabular template. Further, *berth number: two-five* is neither fully specified nor indeed fully specific. Berth 25 in which harbor? In a fully formal system some reference-fixing device, such as described above in §4.3, will be required.

## 7.2 QUESTION

The SeaSpeak QUESTION—*Paisano, this is Shell Southport. QUESTION: What is your ETA at Dover East?*—is straightforwardly represented in a tabular (semistructured) format as follows:

To: Paisano	From: Shell Southport
Question	
ETA	
Subject	Paisano
Place	Dover East
Time	?

The translation to and from FLBC is straightforward. We model questions on the “wife beating” model (*Senator, when did you stop beating your wife?*). That is, questions implicitly state a presumption, then ask the addressee to comment on some aspect of it. In this SeaSpeak example, we interpret Shell Southport to be presuming that Paisano has an ETA at Dover East and then to be asking Paisano to describe the time of that ETA. Paisano’s answer—*Shell Southport, this is Paisano. ANSWER: My ETA at Dover East is: time: one-five-three-zero GMT*—also fits the tabular form easily,

and the translation to FLBC or to an XML semistructured form is straightforward. Here is Paisano’s answer.

To: Shell Southport	From: Paisano
ANSWER	
ETA	
Subject	Paisano
Place	Dover East
Time	(15:30, GMT)

## 7.3 REQUEST

The SeaSpeak message—*Shell Southport. This is Paisano. REQUEST: Please supply bunkers: quantity: two thousand metric tonnes. Over.*—might be tabulated as follows.

Request	
Supply	
Agent	Shell Southport
Benefactive	Paisano
Theme	bunkers
Quantity	2000
Units	metric tonne

Note the potential ambiguity: Are we talking about 2000 bunkers and one supply event or 2000 supply events each supplying one bunker? Surely the former. The correct meaning is easily captured in the logic and the FLBC, and with a slight extension of notation also captured in tabular form:

Request	
Supply	
Agent	Shell Southport
Benefactive	Paisano
Theme	bunkers
bunkers	
Quantity	2000
Units	metric tonne

Shell Southport’s response—*Paisano. This is Shell Southport. REQUEST RECEIVED: Supply bunkers: quantity: two thousand metric tonnes, positive.*—is challenging because *positive* is a subordinating clause. Roughly, the message is *Shell Southport agrees that [subordinating] Shell Southport supplies Paisano. . .* This may be captured in tabular form (and translated without loss to FLBC) as:

Request Received	
Supply	Positive
Agent	Shell Southport
Benefactive	Paisano
Theme	bunkers
bunkers	
Quantity	2000
Units	metric tonne

## 7.4 INSTRUCTION: Giving Reasons

In this message, Paisano acknowledges an instruction, declines to follow it, and gives a reason why: *Shell Southport, this is Paisano. INSTRUCTION-RECEIVED: Stop immediately, negative: reason: I am towing now.* Here is the tabular form, capturing the doubly subordinating message.

Instruction Received	Reason
Stop	Negative
Agent	Paisano
Theme	Paisano
Time	now
Reason	
Tow	
Agent	Paisano
Theme	-
Time	now

## 8. SUMMARY AND DISCUSSION

The present document is something of a position paper. It advocates a vision in the development of computerized support for legal reasoning and it presents evidence indicative of the plausibility of that vision. The pieces of the story might usefully be assembled as follows.

1. Original creation of documents (and records) of legal import in either semistructured or fully formalized format offers the prospect of greatly reducing the cost and expanding the scope of knowledge engineering for legal reasoning.
2. Such origination of legal documents will likely rely on special-purpose sublanguages. If a domain is sufficiently important and well specified to support a sublanguage, even an informal one, that sublanguage becomes a potential target for full or partial formalization. SeaSpeak is one such example. There will be many others.
3. Sublanguages will vary in their expressiveness. and in their completeness along dimensions of lexicon (vocabulary including semantics) and grammar (syntax).
4. A more limited sublanguage, perhaps specified in an XML DTD, will result in a semistructured document.
5. A more complete sublanguage, perhaps specified in in a broadly logical language such as the FLBC discussed above, may result in a fully structured document.
6. We can exploit the structure of documents (essentially, exploit the grammar) in several ways. At a high level, we can do three things: we can build specific data mining models, we can mine the corpus for interesting, unexpected facts and associations, and we can query for explicit items of interest.
7. Tabular message structuring, of the sort described briefly in §7, presents the prospect of reasonable expressive power linguistically in combination with substantial formalization. We entertain hopes that the format will prove usable. If users can learn graffiti for their PDAs, perhaps they may tabular messaging with machines. The tabular format, however, may have difficulty representing universal quantifiers (see example in §4.3). Much investigation remains to be done before these ideas will have been demonstrated to be practicable.

## 9. REFERENCES

- [1] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the web: from relations to semistructured data and XML*. Morgan Kaufmann, San Francisco, CA, 2000.
- [2] T. J. M. Bench-Capon, G. O. Robinson, T. W. Routen, and M. J. Sergot. Logic programming for large scale applications in law: A formalisation of supplementary benefit legislation. In *Proceedings of the first international conference on Artificial intelligence and law*, pages 190–198. ACM Press, 1987.
- [3] J. Benyon. Building police co-operation: The European construction site around the third pillar. World Wide Web file, Accessed 28 February 2003. Dated 1996. <http://www.psa.ac.uk/cps/1996/beny.pdf>.
- [4] C. Cardie. Empirical methods in information extraction. *AI Magazine*, 18(4):65–80, 1997.
- [5] Y. Choueka, M. Cohen, J. Dueck, A. S. Fraenkel, and M. Slag. Full text document retrieval: Hebrew legal texts (report on the first phase of the responsa retrieval project). In *Proceedings of the 1971 international ACM SIGIR conference on Information storage and retrieval*, pages 61–79. ACM Press, 1971.
- [6] A. Daskalopulu and M. J. Sergot. A constraint-driven system for contract assembly. In *Proceedings of the 5th ACM International Conference on AI and Law (ICAIL-95)*, pages 62–70. ACM Press, 1995.
- [7] A. Daskalopulu and M. Sergot. The representation of legal contracts. *AI and Society*, 11(1/2):6–17, 1997.
- [8] D. Davidson. *Essays on Actions and Events*, chapter The Logical Form of Action Sentences, pages 105–148. Clarendon Press, Oxford University Press, Walton Street, Oxford OX2 6DP, United Kingdom, 1980. ISBN: 0-19-824637-4.
- [9] J. P. Dick. Conceptual retrieval and case law. In *Proceedings of the first international conference on Artificial intelligence and law*, pages 106–115. ACM Press, 1987.
- [10] E. Fitzpatrick, J. Bachenko, and D. Hindle. The status of telegraphic sublanguages. In R. Grishman and R. Kittredge, editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pages 39–51. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ, 1986.
- [11] Z. Harris. *Mathematical Structures of Language*. John Wiley & Sons, New York, NY, 1968.
- [12] R. K. Harrison. Bibliography of planned languages (excluding esperanto). World Wide Web, 1992-2002, accessed December 2002. <http://www.invisiblelighthouse.com/langlab/bibliography.html>.
- [13] J. Higginbotham, F. Pianesi, and A. C. Varzi, editors. *Speaking of Events*. Oxford University Press, New York, NY, 2000. ISBN: 0-19-512811-7.
- [14] R. D. Holowczak and N. R. Adam. Information extraction based multiple-category document classification for the global legal information network. In *Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence*, pages 1013–1018. AAAI Press / MIT Press, 1997.
- [15] E. Johnson. LinguaNet final report. World Wide Web file, 1998. [http://www.hltcentral.org/usr\\_docs/project-source/linguanet/Final-Report/Final-Report.html](http://www.hltcentral.org/usr_docs/project-source/linguanet/Final-Report/Final-Report.html), accessed 28 February 2003.
- [16] E. Johnson. Talking across frontiers. In *Proceedings of the International Conference on European Cross-Border Co-operation*, forthcoming. Available at: <http://www.prolingua.co.uk/talking.pdf>. Accessed December 2002.

- [17] E. Johnson et al. *PoliceSpeak: Police Communications and Language, English-French Lexicon*. PoliceSpeak Publications, Cambridge Research Laboratories, 181a Huntingdon Road, Cambridge, CB3 0DJ, 1993. ISBN: 1 898211 01 9.
- [18] D. Jurafsky and J. H. Marton. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, New Jersey, USA, 2000.
- [19] K. Karmiloff and A. Karmiloff-Smith. *Pathways to Language: From Fetus to Adolescent*. Harvard University Press, Cambridge, MA, 2001.
- [20] S. Kerrigan and K. H. Law. Logic-based regulation compliance-assistance. In *Proceedings of ICAIL '03, International Conference on Artificial Intelligence and Law (ICAIL-03)*. ACM Press, 2003.
- [21] S. O. Kimbrough. On representation schemes for promising electronically. *Decision Support Systems*, 6(2):99–122, 1990.
- [22] S. O. Kimbrough. Reasoning about the objects of attitudes and operators: Towards a disquotational theory for representation of propositional content. In *Proceedings of ICAIL '01, International Conference on Artificial Intelligence and Law*, 2001.
- [23] S. O. Kimbrough and S. A. Moore. On automated message processing in electronic commerce and work support systems: Speech act theory and expressive felicity. *ACM Transactions on Information Systems*, 15(4):321–367, October 1997.
- [24] S. O. Kimbrough and Y.-H. Tan. On lean messaging with unfolding and unwrapping for electronic commerce. *International Journal of Electronic Commerce*, 5(1):83–108, 2000.
- [25] A. Large. *The Artificial Language Movement*. Oxford University Press, New York, NY, 1985.
- [26] R. Larson and G. Segal. *Knowledge of Meaning: An Introduction to Semantic Theory*. The MIT Press, Cambridge, Massachusetts, 1995. ISBN: 0-262-62100-2.
- [27] G. Lau, K. H. Law, and G. Wiederhold. Similarity analysis on government regulations. Submitted to *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington DC, Aug 24-27, 2003, Stanford University, 2003. <http://www.stanford.edu/~gorielau/sigkdd.pdf>.
- [28] LinguaNet. The linguaset project. <http://www.cbs.dk/departments/fir/linguaset/>, accessed 2003-02-28.
- [29] T. C. Macaulay. *Interlanguage*. The Clarendon Press, Oxford, UK, 1930.
- [30] A. Marchetti, F. Megale, E. Seta, and F. Vitali. Using xml as a means to access legislative documents: Italian and foreign experiences. *ACM SIGAPP Applied Computing Review*, 10(1):54–62, 2002.
- [31] A. Margelisch. A state of the art report on legal knowledge-based systems. [citeseer.nj.nec.com/187534.html](http://citeseer.nj.nec.com/187534.html), accessed 2003-02-28.
- [32] C. K. Ogden. *Basic English : a general introduction with rules and grammar*. K. Paul, Trench, Trubner, London, UK, 1938.
- [33] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD 1998)*, pages 94–100. ACM Press, August 1998.
- [34] T. Parsons. *Events in the Semantics of English: A Study in Subatomic Semantics*. Current Studies in Linguistics. The MIT Press, Cambridge, MA, 1990. ISBN: 0-262-66093-8.
- [35] ProLingua. Operational communications, controlled languages, computing. World Wide Web page, Accessed 28 February 2003. <http://www.prolingua.co.uk/>.
- [36] R. Quinlan. *R. C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.
- [37] I. A. Richards. *Basic English and Its Uses*. W. W. Norton & Company, Inc., New York, NY, 1943.
- [38] D. E. Rose and R. K. Belew. Legal information retrieval a hybrid approach. In *Proceedings of the second international conference on Artificial intelligence and law*, pages 138–146. ACM Press, 1989.
- [39] N. Sager. Sublanguage: Linguistic phenomenon, computational tool. In R. Grishman and R. Kittredge, editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pages 1–17. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ, 1986.
- [40] E. Schweighofer and D. Merkl. A learning technique for legal document analysis. In *Proceedings of the seventh international conference on Artificial intelligence and law*, pages 156–163. ACM Press, 1999.
- [41] E. Schweighofer, A. Rauber, and M. Dittenbach. Automatic text representation, classification and labeling in European law. In *Proceedings of the 8th international conference on Artificial intelligence and law*, pages 78–87. ACM Press, 2001.
- [42] J. R. Searle and D. Vanderveken. *Foundations of Illocutionary Logic*. Cambridge University Press, Cambridge, England, 1985.
- [43] M. Sergot. The representation of law in computer programs. In T. Bench-Capon, editor, *Knowledge-Based Systems and Legal Applications*, pages 3–67. Academic Press, London, UK, 1991.
- [44] M. J. Sergot, F. Sadri, R. A. Kowalski, F. Kriwaczek, P. Hammond, and H. T. Cory. The British Nationality Act as a logic program. *Communications of the ACM*, 29(5):370–386, 1986.
- [45] C. Swanland. *Basic English for Science & Technology*. Intercultural Press, Chicago, IL, 1980.
- [46] F. Weeks, A. Glover, E. Johnson, and P. Strevens. *Seaspeak Training Manual: Essential English for International Maritime Use*. Pergamon Press, Oxford, UK, 1988. Available via <http://www.maritimeusa.com>.

File: kimbrough.tex/pdf née:

\$Id: gen-struct-legal-docs.tex,v 2.3 2003/05/19\$