

# On Original Generation of Structure in Legal Documents\*

Steven O. Kimbrough  
Thomas Y. Lee  
Balaji Padmanabhan  
Yinghui Yang  
University of Pennsylvania  
3730 Walnut Street  
Philadelphia, PA 19104-6340

June 25, 2003

---

\*File: icaill-2003-foils.pdf.

# Means of Obtaining Structured Information from Documents

1. Manually symbolize
  2. Automatically symbolize (e.g., via parsing)
  3. Accept minimally structured documents (e.g., use IR or IE techniques)
  4. Create the documents in structured form
- . . . plus mixtures.

## **Idea: Originate legal documents in structured form**

NB: Not our original idea. Wish to explore it. Outline:

- Degrees and kinds of structuring
- Where might the structuring be used?
- What might be done with documents so structured?
- Will it work?

NB: This is really a position paper or thought piece. Will present some results, but they are hardly completed.

## Degrees and kinds of structuring

- Fully formalized. Example: FOL and beyond.
- Semi-structured. Example: XML with carefully articulated DTDs.
- Quasi-structured. Examples: Recipes, weather reports, personals ads.
- Lexically restricted. (Unstructured but with a characteristic lexicon and semantic relationships.) Example: Scientific abstracts in a specialized subfield; similarly legal abstracts.

# Observation

In many areas of activity, specialized sublanguages are in use.

- Sublanguage (aka: microlanguage): “A sublanguage is characterized by distinctive specializations of syntax and the occurrence of domain-specific word subclasses in particular syntactic combinations.” (Sager). Examples of broadly legal import: SeaSpeak, AirSpeak, PoliceSpeak (and much else).
- Arise naturally, by design or as an unintended side-effect of practice.
- People train in them.

## Comments on Sublanguages

- Found in the literature of academic specialties, e.g., peptide catalysis.
- Related concept: telegraphic languages. Examples: submarine communication, stock trading, etc. “Work request submitted.” “No parts required.” “72 man hours expended.”
- Common themes: simplification and specialization. See C.K. Ogden and Basic English. Eliminate “climb” in favor of “go up.”
- Likely to be found (or could be imposed?) in legal decisions, regulatory documents, etc.

## Suggestion (Position of the Paper)

Sublanguages—arising spontaneously or explicitly designed—will often be felicitous targets of formalization.

- Formalize a sublanguage and use the formalized version to create new documents/records.
- Formalize a sublanguage and automatically translate to it from natural language texts.

Focusing on the former, wish to present some evidence supporting the suggestion.

NB: KISS principle (“leanness”). Focus on tabular format for user interface.

Example Sublanguage:  
SeaSpeak

# SeaSpeak

- Quasi-structured sublanguage in use for maritime communications. Accepted as an international standard.
  - Conversation structure: simple; skip in interests of time.
  - Message structure: F(C)
    - F: Illocutionary force
    - C: Content
- Warms the heart of a speech act theorist.

## F: Illocutionary force structure

SeaSpeak has seven message markers (each with a mirroring reply-marker):

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)

## F: Illocutionary force structure: Interpretation

$F(C) \rightsquigarrow$  *Speaker s F's addressee a that C.* (Subordinating construction: *F that C.*)

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...*

## **C: Message content structure. Example messages.**

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

## C: Message content structure

- Contents governed by message markers are quite simple—1 or 2 sentences and highly constrained and stylized.
- Message markers are not iterated.
- With some exceptions, content sentences are simple, not complex: no embedded clauses or subordinators (e.g., *that P* constructions).

But note: *authorise* (to *authorise*). The INTENTION message marker: speaker may tell addressee the intentions of a third party: *The icebreaker intends to assemble the convoy at time: zero-five-three-zero GMT.*

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

Honoring requests:

*Shell Southport. This is Paisano. REQUEST: Please supply bunkers: quantity: two thousand metric tonnes. Over.*

*Paisano. This is Shell Southport. REQUEST RECEIVED: Supply bunkers: quantity: two thousand metric tonnes, positive.*

## C: Message content structure

- Normative elements? E.g.,  $C \stackrel{?}{=} a \text{ ought to see to it that } C'$ . Not (so far as we can see) in the messages. Rather, implied by certain messages under the conditions of discourse.

*INFORMATION: You are fishing in a prohibited area.*

## **Additional C structure: noun sets**

General, proper nouns:

- Ships: Kalong Treasure, Atlantic Rover, Silja Queen, Paisano
- Aircraft: Watchdog Three
- Places/entities: Land's End, Southport Harbour, Southport Operations, Falmouth Coastguard, Goteborg Port Traffic, Shell Southport

## **Additional C structure: noun sets, explicit**

- Ship, boat and aircraft types: barge, cable ship, collier, corvette, derelict, ferry, frigate, yacht, seaplane. . .
- On-board terminology (parts and equipment): gangway, anchor, bitt, boarding ladder, boat, bollard, bow, derrick, fuel, gear, life raft, light, loran, rocket, rudder, stern, bilge, ballast, winch. . .
- Engineering: aerial, air filter, alarm, alternator, ammeter, amplifier, armature, auxiliary steering gear. . .
- Safety, navigation and pilotage: accident, breakers, . . .

- Business and miscellaneous: agent, agreement, competent authority, de-rat certificate, dock regulations, doctor, drinking water, duty, officer in charge, passenger, pollution
- Buoys, lights and beacons: beacon, buoy, character (lighted aid to navigation), composite group occulting, continuous quick, radar, traffic control signals, continuous very quick. . .
- Post and coast features and installations: anchorage, archipelago, bank, bar, bay, beach, bend, bert, boat, boom, bottom, bridge,. . .

## **Additional C structure: locatives, explicit**

- On-board terminology (general): abaft, abeam, abreast, aboard, aft, after, ahead, alongside, amidships, athwart, . . .
- Safety, navigation and pilotage: close to the surface, closest point of approach (CPA), dangerous quadrant

## **Additional C structure: verbs, explicit**

- Safety, navigation and pilotage: awash, aweigh, beneaped, berth (to berth), blocked, out-of-control,. . .
- Business and miscellaneous: de-rat (to de-rat). . .

# Tabular : FLBC & SeaSpeak

Recall:

- High-level structure:  $F(C)$ .  $F$ : illocutionary force.  $C$ : propositional content
- $F$  in SeaSpeak: one of 7 (+7) message markers
- $F$  structure:  $\langle message-marker \rangle(e) \wedge Speaker(e, s_1) \wedge Content(e, [C]) \wedge Addressee(e, a_1) \wedge Cul(e, now)$

$s_1$	$\langle message-marker \rangle$	$a_1$	$now$
$C$			

## *C* Structure

- *C*: normally a simple sentence or two
- *C* ::= CSF  
(core sentence frame = verb + essential thematic roles)

# Thematic Roles

Role	Interpretation; typical use
<i>Agent</i> ( $e, x$ )	$x$ initiates event $e$ ; subject of sentence, Intentional/volitional causer of event
<i>Force</i> ( $e, x$ )	$x$ initiates event $e$ ; subject of sentence, Non-agent causer of event
<i>Content</i> ( $e, [p]$ )	that $p$ is the propositional content of $e$
<i>Theme</i> ( $e, x$ )	$e$ does something to $x$ ; direct object of sentence
<i>Source</i> ( $e, x$ )	$e$ is directed from $x$ ; often, indirect object of sentence
<i>Goal</i> ( $e, x$ )	$e$ is directed at $x$ ; often, indirect object of sentence
<i>Location</i> ( $e, x$ )	$e$ is situated at $x$ ; subsumes many prepositions
<i>Benefactive</i> ( $e, x$ )	$e$ is for the sake of $x$ ; indirect object aka: <i>Beneficiary</i>
<i>Sake</i> ( $e, x$ )	$e$ is for the sake of $x$ ; fixes reference to previous message
<i>Instrument</i> ( $e, x$ )	secondary cause of the event; object of “with” aka: <i>Performer</i>
<i>Experiencer</i> ( $e, x$ )	individual (person) experiencing $e$ ; subject of sentence
<i>Cul</i> ( $e, t$ )	event $e$ culminates at, or during, time $t$
<i>Hold</i> ( $e, t$ )	process or state $e$ holds or obtains at, or during, time $t$

Table 1: Basic, standard thematic roles

## Example CSFs: go

go	
Experiencer	
Source	
Goal	
Cul Hold	

Template CSF.



go	
Experiencer	Paisano
Source	Dover East
Goal	-
Cul	tomorrow

Instance CSF. "Paisano will leave Dover East tomorrow."

## go Continued

go	
Experiencer	Paisano
Source	–
Goal	Dover East
Cul	tomorrow

Instance CSF. “Paisano will arrive at Dover East tomorrow.”

go	
Experiencer	Paisano
Source	Dover East
Goal	Boston
Hold	tomorrow

Instance CSF. “Paisano will be going from Dover East to Boston tomorrow.” NB: Cul  $\rightsquigarrow$  “Paisano will go from Dover East to Boston tomorrow (leave and arrive).” [Comc]

## More Examples of CSFs

beneap	
Experiencer	
Location	
Cul Hold	

Template CSF.  
[Instrument=the  
beneap tide]

hoist	
Agent	
Theme	
Source	
Goal	
Cul	

Template CSF.  
[Hold]

heave-away	
Agent	
Source	
Cul	

Template CSF.  
[Hold; Agent-  
Theme]

## More Examples of CSFs

circulate	
Agent	
Location	
Cul Hold	

Template CSF.  
[Agent-Theme  
vs Agent &  
Theme]

overtake	
Agent	
Theme	
Location	
Cul	

Template CSF.  
[Hold]

approach	
Experiencer	
Goal	
Cul	

Template CSF.  
[Hold]

## Locators: Additional, Optional Thematic Roles

- TLOCs (temporal locators): *Cul, Hold, Comc*
- GLOCs (general/geographic locators): *Location of, up, down, towards, away-from, left-of, front-of, near to, far from, north of, etc.*
- $C ::= \text{CSF} | \text{CSF TLOC} | \text{CSF GLOC}$   
... but see RMeS below.

Also: *Location-on* measured by, e.g., *top, bottom, side, 11 o'clock, etc.*

## Measurers: Additional, Optional Thematic Roles

- Measurers:  $Measure(\langle thing \rangle, [\langle attribute \rangle], \langle scale \rangle, \langle value \rangle)$

“ $x$  is brown”  $\rightsquigarrow Measure(x, color, brown)$

“ $x$  is large”  $\rightsquigarrow Measure(x, size(x), large)$  (Small elephants and large mice.)

“ $e$  is violent”  $\rightsquigarrow Measure(e, violence, high)$  or  $Measure(e, violence(e), high)$  (Violent stabbing vs. violent argument)

“ $x$  is 17 meters long”  $\rightsquigarrow Measure(x, length, meter, 17)$

## Tabular Form

MM	⟨speaker⟩	⟨message-type⟩	⟨addressee⟩	⟨Cul⟩
CSF	⟨verb⟩	[⟨denial⟩]		
	⟨Theta role⟩	[⟨value⟩]		
	⋮	⋮		
	⟨Theta role⟩	[⟨value⟩]		
TLOC	Cul Hold	[⟨value⟩]		
GLOC	⟨thing⟩	⟨value⟩		
Mes	⟨thing⟩	[⟨attribute⟩]	⟨scale⟩	⟨value⟩
Mes	⋮	⋮	⋮	⋮
Mes	⟨thing⟩	[⟨attribute⟩]	⟨scale⟩	⟨value⟩

MM=Message Marker. CSF= Core Sentence Frame. TLOC=Temporal Locator. GLOC=General Locator. Mes=Measurement.

# Tabular Form: Template

MM				
CSF				
TLOC				
GLOC				
Mes				
Mes				
Mes				

## Example (p. 97 SeaSpeak manual)

“REQUEST: Please send, quantity: five acetylene cylinders.”

MM	$s_1$	Request	$a_1$	now
CSF	send			
	Agent	$a_1$		
	Benefactive	$s_1$		
	Theme	acetylene cylinder		
TLOC	Cul	soon		
Mes	acetylene cyliner	quantity	unit	5

Note: Benefactive could be left unspecified, with the assumption that the speaker,  $s_1$ , is the benefactive.

## Example (p. 97 SeaSpeak manual)

“INFORMATION: The pilot is waiting now, position: near buoy number: two-six.”

MM	$s_1$	Inform	$a_1$	now
CSF	wait			
	Agent	the-pilot		
	Theme	–		
TLOC	Hold	now		
GLOC	wait	place1		
RMes	(place1, buoy)	distance	dummy	near
Mes	buoy	ID	number	26

RMes: relational measure. “dummy” for nominal scale. “the-pilot” a matter for reference fixing. Theme is null because what the wait is for is unspecified (and may not exist).

## Example: What is your ETA at Dover East?

- *ANSWER: My ETA at Dover East is: time: one-five-three-zero GMT.*
- Modeling/semantic issue: Is this about a future arrival event whose culmination is estimated? or is this about an object, an ETA, with a characteristic property or is this about an ETA event that is said to happen in the future? We take the latter option.
- Stylistic variant:  *$e_1$  is an ETA event. It will be at/to Dover East at time 15:30 GMT. All this will happen to the speaker,  $s_1$ .*
- In FOL:  $ETA(e_1) \wedge Experiencer(e_1, s_p) \wedge Goal(e_1, 'Dover East') \wedge Cul(e_1, 15 : 30, GMT)$

## Comments on Example

- *Experiencer*, *Goal*, *Cul* fit naturally with *ETA* as a verb/event in this domain. This affords semi-structuring, e.g., supposing the addressee is the Paisano.

Question	
ETA	
Subject	Paisano
Place	Dover East
Time	?

Answer	
ETA	
Subject	Paisano
Place	Dover East
Time	15:30 GMT

Note:  $\approx$  Questions make presuppositions and then request the addressee to describe some aspect of what is presumed. Or, all questions are wife-beating questions.

## Comments on Example

Note generalizations, e.g.

- Who is arriving at Dover East at 15:30? Where are you arriving at 15:30?

Question	
ETA	
Subject	?
Place	Dover East
Time	15:30 GMT

Question	
ETA	
Subject	Paisano
Place	?
Time	15:30 GMT

## Example: Honoring requests

- Recall:

*Shell Southport. This is Paisano. REQUEST: Please supply bunkers: quantity: two thousand metric tonnes. Over.*

*Paisano. This is Shell Southport. REQUEST RECEIVED: Supply bunkers: quantity: two thousand metric tonnes, positive.*

- Note: *positive* is a subordinating clause  $\approx$  *Shell Southport agrees that Shell Southport supplies Paisano...*

## Example, semi-structured

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

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

Comment: The resources of FLBC are ample for handling the underlying logic.

## Example: Giving reasons

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

Instruction	
Stop	
Agent	Paisano
Time	now

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

# What might be done with documents so structured?

Broadly

1. Automation
2. Queries, information extraction
3. Discovery

# Automation

Communication with a machine at (at least) one end.

- A means of creating/generating original documents
- Language translation

# Figure 1

REQUEST: Please send, quantity: five acetylene cylinders.

MM	$s_1$	REQUEST	$a_1$	now
CSF	send			
	Agent	$a_1$		
	Benefactive	$s_1$		
	Theme	acetylene cylinder		
TLOC	Cul	soon		
Mes	acetylene cylinder	quantity	unit	5

语气	发言人	请求	受言人	现在
句子主干	送			
	主语	受言人		
	受益人	发言人		
	直接宾语	乙炔汽缸		
时间	发生	马上		
度量	乙炔汽缸	数量	单位	5

## Figure 2

Axel. This is Northport Pilot.  
 ADVICE: Alter course to port.

MM	Northport Harbor	ADVICE	Axel	now
CSF	alter-course			
	Agent	Axel		
	Goal	port		
TLOC	Cult	soon		

语气	Northport Harbor	建议	Axel	现在
句子主干	改变航向			
	主语	Axel		
	目标	港口		
时间	发生	马上		

## Will it work? Summary.

- Potential benefits of original generation of legal documents.
- How? Details?
- Exploit sublanguages, even create them where needed.
- Tabular structure, drawing on FLBC experience, appears promising for representation of sublanguage expressions.
- Given tabular structure (or something like it) for a workable sublanguage, potential benefits appear realizable.

## Will it work? Heavy simplification with the tabular format.

- Logical constants limited to  $\wedge$  and  $\neg$ .
- No quantification. But consider: “No ships are in the harbor.”

Empty	
In	the-harbor
Theme	ships
Time	now

Recall: elimination of “climb” in favor of “go up.” Also, note: ETA.

## Will it work? Some Envisioned Uses.

- Support for language translation. (Table  $\rightsquigarrow$  output.)
- Support for message clarification. (Table  $\rightsquigarrow$  output.)
- Historical reconstruction of events. (Table  $\rightsquigarrow$  output.)
- Evidence exploitation, prediction, event analysis. (Table  $\rightsquigarrow$  output.)
- Specification of rules, terms and conditions, etc. (Sublanguage  $\rightsquigarrow$  Table?)
- Compliance checking. (Sublanguage  $\rightsquigarrow$  Table?)

## Will it work? Conclusion.

- The tabular structure, sketched above, appears to work well in SeaSpeak and in other sublanguage contexts. “to work well” in the sense of “represent the meaning reasonably well.” Viz., our slogging through SeaSpeak.
- The tabular structure seems to have strong benefits as a user interface concept. Viz., Chinese translation; what we know about similar interfaces, e.g., QBE.
- The tabular structure seems to be an apt target for the output of parsing an unformalized sublanguage, e.g., design specifications in the sublanguage, parse, and generate tabular expressions.
- The tabular structure would seem to afford good support for automated querying of documents. Viz., semi-structured query languages.

All this remains to be tested much more rigorously.

\$Id: icail-2003-foils.tex,v 1.5 2003/06/13 21:26:20 sok Exp \$