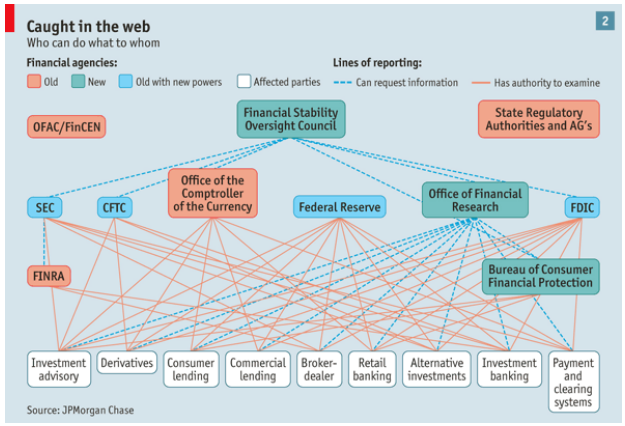
Knowledge Extraction, Representation and Reasoning Research Prospectus

From Linguistics to Knowledge

Vijay Saraswat

The Regulatory Compliance problem



Understand the regulatory constraints imposed by the world on the enterprise

What are all the regulations that impact my area of responsibility?

How has the regulation changed? How does it affect my existing controls, policies, procedures?

What does this reg mean to my <X> line of business?

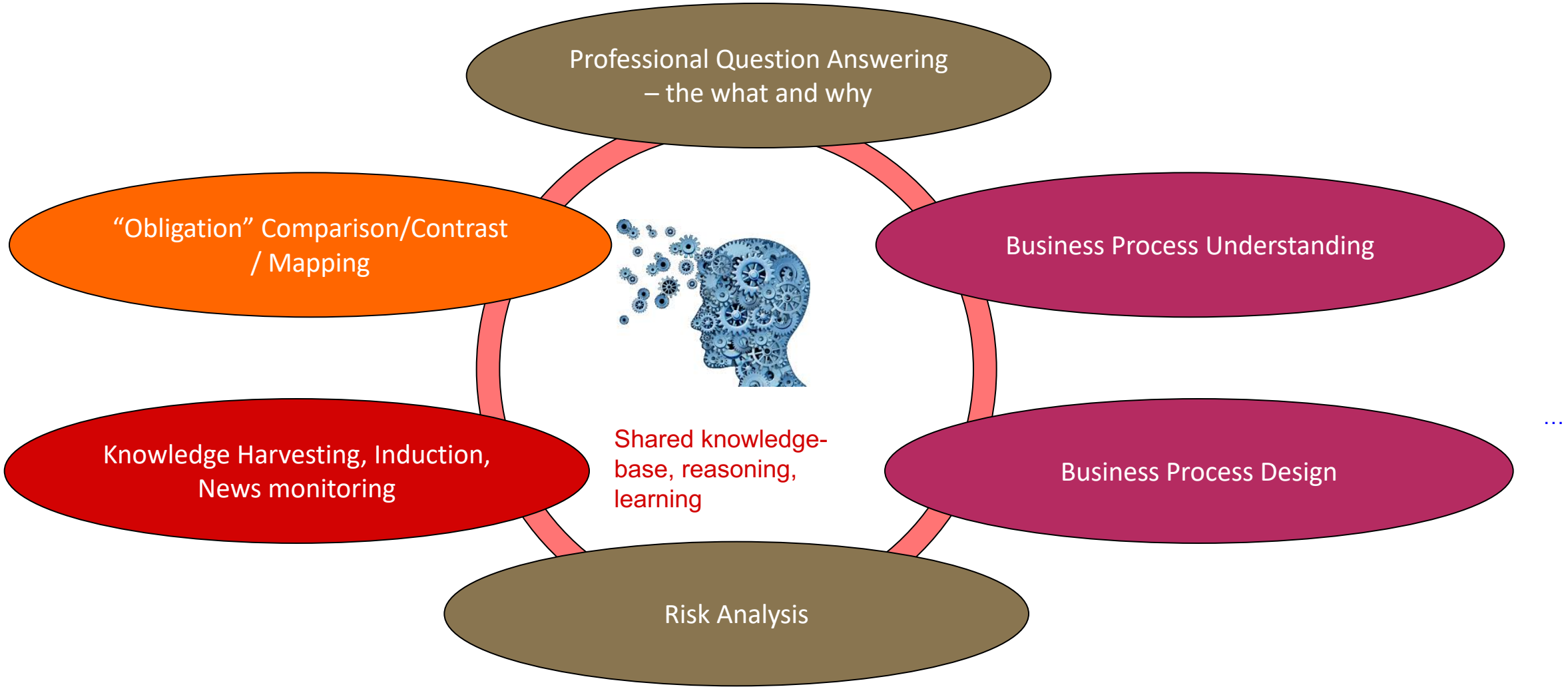
My enterprise wants to enter into business <Y> in country <Z>. What regulations apply?

Understand the enterprise

Full traceability from regulation to products, policies, risks and controls

Do this continuously as regulations, people, systems change.

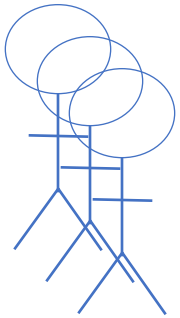
The Research Landscape for Compliance



...

Deep understanding of domain Deep understanding of business processes

Generic Upton Architecture



Interaction/User Experience / Workflow

SAW
(integrated dialog, NLG + UX for various application operations, workflow)

Quasi-logical forms

Task input

Problem-solvers/Reasoners

Directed Reasoning
Contextual Reasoning
Textual Reasoning
Explanation Generation

Machine learning

Task Descriptions

Knowledge Store & DB

Scaled-out representations

Task output

Auxiliary sources (e.g. answer store, quote generator)

Parsing
Corpus creation
Sentence operations
Linguistics to Knowledge

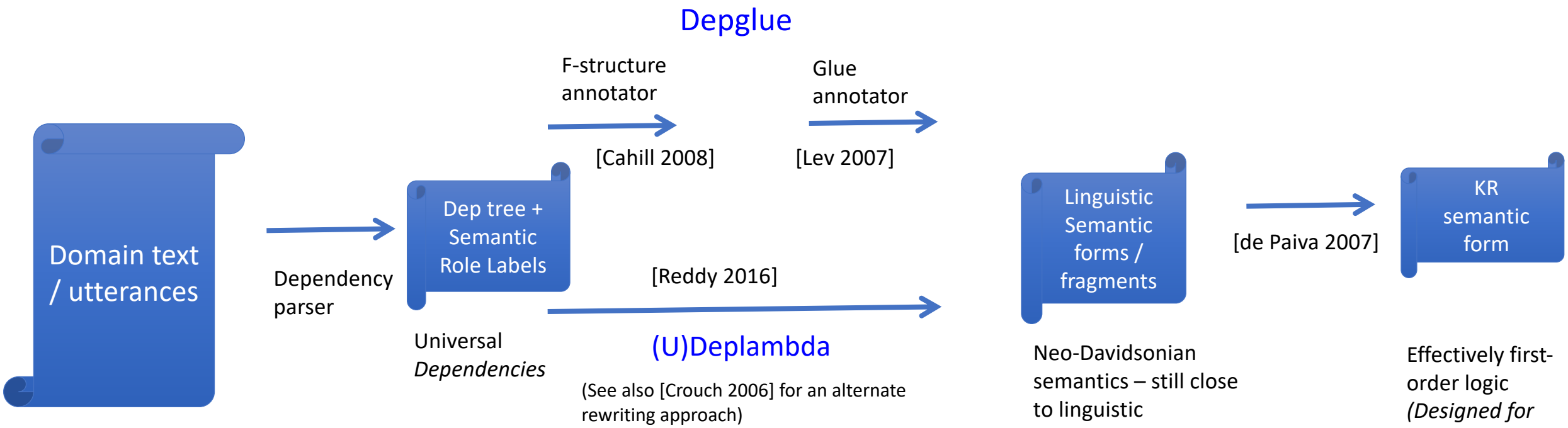
Machine learning

Knowledge Extractor

Continuous Ingestion of documents

Content Acquirer

From linguistics to knowledge



Using unsupervised/supervised techniques, symbolic / neural techniques.
Note: Semantic Parsers typically require (utterance, semantic)* corpus. *Need to develop sembanks for our discourse domains*

Issues to consider:

- Discourse (cf glue for DRT)
- Treatment of ambiguity (packed representation?)

Compliance | Legal | Financial | Contracts …

Note: PARC, Stanford researchers (Bobrow, de Paiva, Crouch, Karttunen, …) pioneered this line of attack (DARPA AQUAINT c 2007)

Example Task: Professional Question Answering

Article 4: 1. Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of:

(a) all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

(b) portable batteries or accumulators, including those incorporated into appliances, that contain more than 0,002 % of cadmium by weight.

2. The prohibition set out in paragraph 1(a) shall not apply to button cells with a mercury content of no more than 2 % by weight.

DIRECTIVE 2006/66/EC

Q: Can I use battery M6512 that contains 0.15% mercury by weight for a hand watch, after 2015, in Europe?

A: If it is a button cell. (Or, possibly, in some exceptional cases.)

Labeling structure – numbered articles, branches.

References to labels.

Unusual "entities": Directive 2000/53/EC, 'Member States', generic elements (mercury, cadmium, batteries, accumulators, appliances)

Single sentence may be split across multiple paragraphs, with multiply nested bullets

Complex (vs Simple / Compound) sentences – may have a large number of clauses

Modification

Coordination – batteries or accumulators

Negation – whether or not

Anaphora -- which

Long-distance dependencies -- relative clauses:
whether or not incorporated into appliances
That contain more than … by weight

Quantifiers:
All batteries or accumulators…

Example Task: Professional Question Answering

DIRECTIVE 2006/66/EC

DIRECTIVE 2006/66/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 6 September 2006

on batteries and accumulators and waste batteries and accumulators and repealing Directive 91/157/EEC

(Text with EEA relevance)

Preamble

Definitions

Complex references

3. The prohibition set out in paragraph 1(b) shall not apply to portable batteries and accumulators intended for use in:
(a) emergency and alarm systems, including emergency lighting;
(b) medical equipment; or
(c) cordless power tools.
4. The Commission shall review the exemption referred to in paragraph 3(c) and submit a report to the European Parliament and to the Council by 26 September 2010, together, if appropriate, with relevant proposals, with a view to the prohibition of cadmium in batteries and accumulators.

The Real problem is more complex!

Example Task: Professional Question Answering

DIRECTIVE 2013/56/EU

Article 1
Directive 2006/66/EC is amended as follows:
(1) Article 4 is amended as follows:
(a) paragraph 2 is replaced by the following:
'2. The prohibition set out in paragraph 1(a) shall not apply to button cells with a mercury content of no more than 2 % by weight until 1 October 2015.';
…

Subsequent regulations invalidate some portions of old regulation.

Regulation does this by simply replacing clauses in old regulations with new clauses.

Statistical techniques do not understand "quotes"!
Logical techniques can.'

Task: Construct compositionally* the currently active obligations

* Compositionality: Conjunction of obligations extracted separately from Reg A and Reg B should give the right result even if Reg B rupdates Reg A.

Example Task: Professional Question Answering

Article 4: 1. Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of:
(a) all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;
…
2. The prohibition set out in paragraph 1(a) shall not apply to button cells with a mercury content of no more than 2 % by weight.

R: '_ prohibits the placing on the market of _'(State, Item):-
R=rule('Directive 2006/66/EC', ['Article 4', 1, a]),
'member state'(eu,State),
'battery or accumulator'(Item),
applicable(R,Item),
'mercury content'(Item,'by weight',X percent),
{X > 0.0005}.

DIRECTIVE 2006/66/EC

Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

Member States shall prohibit the placing on the market of all batteries or accumulators, that contain more than 0,0005% of mercury by weight.

Q: Can I use battery M6512 that contains 0.15% mercury by weight for a hand watch, after 2015, in Europe?

A: If it is a button cell. (Or, possibly, in some exceptional cases.)

Regulatory Documents have a lot of exploitable structure

DIRECTIVE 2014/65/EU [MiFiD 2]

Scope

Exemptions

Optional Exemptions

Typically, scope specifies which entities are governed by the regulation.

Exemptions carves out exceptions – these can be specific and detailed

DIRECTIVE 2014/65/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 15 May 2014

on markets in financial instruments and amending Directive 2002/92/EC and Directive 2011/61/EU

(recast)

Preamble: 15p, 170 $para_1$

Title I: Scope and Definitions (p26-p40)
Article 1: Scope: 1p, 7 $para_1$, 12 $para_2$
Article 2: Exemptions: 3p, 4 $para_1$, 22 $para_2$
Article 3: Optional Exemptions (Member States choice): 2p, 5 $para_1$, 8 $para_2$
Article 4: Definitions: 8.5p, 63 $para_1$

Title II: Authorisation and Operating Conditions for Investment Firms (p40-79)

Title III: Regulated Markets (Article 44 – 56, p79-92)

Title IV: Position limits and position management controls in … (Articles 57-58, p 92-97)

Title V: Data reporting Services (p 97-105)

Title VI: Competent Authorities (p105-125)

Title VIII: Delegated Acts (Article 89, p125-126)

Final Provisions (Article 90 – Article 97, p126-132)
Annex I-III (p 132-148)

Definitions

Definitions specify meanings for specific terms, sometimes have very complex structure

Sometimes, scope forward references exemptions and overrides them!

Obligations are context dependent

Article 1
Scope
1. This Directive shall apply to investment firms, market operators, data reporting services providers, and third-country firms providing investment services or performing investment activities through the establishment of a branch in the Union.

Task: Develop tests to determine whether credit institutions are authorized by given directive

3. The following provisions shall also apply to credit institutions authorised under Directive 2013/36/EU, when providing one or more investment services and/or performing investment activities:
(a) Article 2(2), Article 9(3) and Articles 14 and 16 to 20,
(b) Chapter II of Title II excluding second subparagraph of Article 29(2),
(c) Chapter III of Title II excluding Article 34(2) and (3) and Article 35(2) to (6) and (9),
(d) Articles 67 to 75 and Articles 80, 85 and 86.

Task: Develop function test. Is credit institution providing one or more investment services? Performing investment activities?

Task: Construct referenced spans

Q: What are the obligations for credit institutions?

Task: Extract obligations only in referenced spans

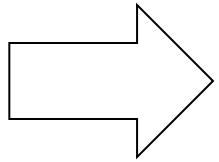
Clean, Atomic Obligation Extraction

Vendor's Program must, at a minimum, contain the following general information security controls:

a. Physical security standards protecting information from unauthorized physical and electronic access.

b. System hardening standards such that default configurations and deployments are appropriately modified to protect XXX Confidential Data. Software must be centrally managed to ensure that unauthorized modifications cannot be made or go undetected.

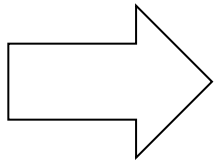
c…



1. Vendor's Program must, at a minimum, contain physical security standards protecting information from unauthorized physical and electronic access.
2. …
3. Vendor's Program must, at a minimum, specify that software must be centrally managed to ensure that unauthorized modifications cannot be made or go undetected.

A financial institution shall collect data regarding applications for, and originations and purchases of, home purchase loans, home improvement loans, and refinancings for each calendar year.

(HMDA)



9 (3 * 3) atomic obligations

1. A financial institution shall collect data regarding applications for home purchase loans for each calendar year.
2. A financial institution shall collect data regarding applications for home improvement loans for each calendar year.

Rewrite original text, keeping provenance, to facilitate downstream processing

Contextual Obligation Extraction

§ 1003.2 Definitions.

Financial institution means:
(1) A bank, savings association, or credit union that:

Home improvement loan means:
(1) A loan secured by a lien on a dwelling that is for the purpose, in whole or in part, of

Application. (1) *In general.* Application means an oral or written request for a home purchase loan, a home improvement loan, or a refinancing that is made in accordance with procedures used by a financial institution for the type of credit requested.

A financial institution shall collect data regarding applications for, and originations and purchases of, home purchase loans, home improvement loans, and refinancings for each calendar year.

Home purchase loan means a loan secured by and made for the purpose of purchasing a dwelling.

Refinancing means a new obligation that satisfies and replaces an existing obligation by the same

definiendum

definiens

mentions

Extract contextual information for obligations: named entities (agencies, acts,…), abbreviations, authority, scope, purpose, definitions (and their use), links, examples, commentary, …

Extract Logical Structure of Obligations

An institution is required to collect data regarding requests under a preapproval program (as defined in § 1003.2) only if the preapproval request is denied or results in the origination of a home purchase loan.

Frame	Obligation
Pattern	[*subject*]{ **is required to** [*act*]}[*object*]{**only if** [*condition*]}
Trace	§ 1003.4(a) Lines 3:5
Slots	**Values**
subject	An institution
act	**is required to** collect
object	data regarding requests under a preapproval program (as defined in § 1003.2)
condition	**only if** the preapproval request is denied or results in the origination of a home purchase loan

Extract logical structure: who shall do what to whom, under what conditions, …➔ Semantic comparison of obligations

Linguistic phenomena in contracts

The effective date of this order will be the later of the Effective Date noted herein or the date User Id is issued .

Customer shall not share its user ID 's and passwords outside the U.S. , nor may it share information accessed under this Order with persons located outside the U.S .

Customers may only make Services under this Order available to entities located in the United States that are subsidiaries , divisions or affiliates , wholly-owned or controlled by Customer (`` US Affiliates '') and identified on a `` Schedule of Affiliates '' attached to this Order and that are not currently eligible to receive any Services included herein under an existing agreement with XXX to support their respective US businesses .

Long-distance dependencies: relative clauses

- that are not currently eligible to receive any Services included herein under an existing agreement with XXX to support their respective US businesses

Complex (vs Simple / Compound) sentences – typically less than ten clauses.

Modification - effective date

Coordination –

- user ID's and passwords
- the later of the Effective Date noted herein or the date User Id is issued
- not share its user ID 's and passwords outside the U.S. , nor may it share information accessed under this Order with persons located outside the U.S

Anaphora -- its user ID's and passwords

Linguistic phenomena in regulations

v) Domestic scheduled commercial banks (other than RRBs) are permitted to open branches, Administrative offices, Central Processing Centres (CPCs) and Service branches in Tier 2 to Tier 6 centres (with population up to 99,999 as per Census 2001 - details of classification of centres tier-wise furnished in Annex 5) and in rural, semi-urban and urban centres in North Eastern States and Sikkim, and to open mobile branches in Tier 3 to Tier 6 centres (with population up to 49,999 as per Census 2001) and in rural, semi-urban and urban centres in North Eastern States and Sikkim without permission from Reserve Bank of India in each case, subject to reporting.

T*arget formalization – currently manually generated*.

Single sentence may be split across multiple paragraphs, with multiply nested bullets

Complex (vs Simple / Compound) sentences – may have a large number of clauses

Modification

Coordination – sentential, NP-, VP-, involving unsaturated predicates

Anaphora -- which

Long-distance dependencies: e.g. relative clauses, wh-questions

Quantifiers

…

Complex and extremely challenging

Linguistic phenomena in prospectuses

1) a) The Fund may exclusively invest in:
…
vii) Money market instruments other than those dealt in on a Regulated Market, if the issue or the issuer of such instruments are themselves regulated for the purpose of protecting investors and savings, and provided that such instruments are:
a.issued or guaranteed by a central, regional or local authority or by a central bank of an EU Member State, the European Central Bank, the EU or the European Investment Bank, a non-EU Member State or, in case of a Federal State, by one of the members making up the federation, or by a public international body to which one or more EU Member States belong; or
b.issued by an undertaking, any securities of which are dealt in on Regulated Markets referred to in 1) a) i) and ii) above; or
c.issued or guaranteed by a credit institution subject to prudential supervision in accordance with criteria defined by European law or by a credit institution which is subject to and complies with prudential rules considered by the CSSF to be at least as stringent as those laid down by the European law; or
d.issued by other bodies belonging to the categories approved by the CSSF provided that investments in such instruments are subject to investor protection equivalent to that laid down in a. b. or c. above and provided that the issuer is a company whose capital and reserves amount to at least ten million Euro (EUR 10,000,000) and which presents and publishes its annual accounts in accordance with the fourth Directive 78/660/EEC, is an entity which, within a group of companies, is dedicated to the financing of the group or is an entity which is dedicated to the financing of securitisation vehicles which benefit from a banking liquidity line.

Single sentence may be split across multiple paragraphs, with multiply nested bullets

Complex (vs Simple / Compound) sentences – may have a large number of clauses

Modification

Coordination – sentential, NP-, VP-, involving unsaturated predicates

Anaphora -- which

Long-distance dependencies: e.g. relative clauses, wh-questions

Quantifiers

…

Complex and extremely challenging

Corporate Loan Agreement

Examples taken from SEC example

Corporate Loan Agreements:Definitions

- "Account Control Agreement": means the Account Control Agreement, dated as of even date herewith, by and among Borrower, Lender and the Account Control Bank, substantially in the form of Exhibit B attached hereto, which grants Lender a Lien on and security interest in the Borrower Account.
- "Account Control Bank": is defined in Section 2.6(b).
- "Additional Advance Date": means, with respect to each Advance subsequent to the initial Advance, the first date upon which all of the conditions precedent set forth in Section 3.2 are satisfied or waived by Lender in its sole discretion. If all Advances are made on the same date, the Additional Advance Date shall be such date.
- "Advance": is defined in Section 2.1.
- "Advance Request": means a notice of advance request in the form of Exhibit D attached hereto, delivered by Borrower to Lender at least two (2) Business Days prior to the Financial Closing Date or an Additional Advance Date, stating the name of the Project in respect of which the Advance is sought and the amount of the Advance requested, the date on which the Advance is to be paid and the account to which the Advance is to be paid.

Extensive list of definitions – 158 running over 20 out of 61 pages.

Corporate Loan Agreements

"Environmental Laws": means any constitutional provision, statute, act, code, law (including common laws), regulation, rule, ordinance, order, decree, ruling, proclamation, resolution, judgment, decision, declaration, or interpretative or advisory opinion or letter of a Governmental Authority pertaining to human health and safety and the environment, in any and all jurisdictions in which the Project Companies are conducting business, or where any real property of the Project Companies is located or where any Hazardous Substances generated by or disposed of by any Project Company are located, including, without limitation, the Clean Air Act, as amended, the Comprehensive Environmental, Response, Compensation, and Liability Act of 1980 ("CERCLA"), as amended, the Federal Water Pollution Control Act, as amended, the Resource Conservation and Recovery Act of 1976 ("RCRA"), as amended, the Safe Drinking Water Act, as amended, the Toxic Substances Control Act, as amended, the Superfund Amendments and Reauthorization Act of 1986, as amended, and any other federal, state, regional or local environmental conservation or protection laws as each may from time to time be amended or supplemented.

Complex clausal structure!

Corporate Loan Agreements

"Event of Bankruptcy": means, with respect to any Person, the occurrence of any of the following events:

(i) the commencement by such Person of a voluntary case concerning itself under the Bankruptcy Code or similar Applicable Law;

(ii) an involuntary case is commenced against such Person and the petition is not controverted within ten (10) days, or is not dismissed within sixty (60) days, after commencement of the case;

(iii) a custodian (as defined in the Bankruptcy Code) is appointed for, or takes charge of, all or substantially all of the property of such Person or such Person commences any other proceedings under any reorganization, arrangement, adjustment of debt, relief of debtors, dissolution, insolvency or liquidation or similar Applicable Law of any jurisdiction whether now or hereafter in effect relating to such Person or there is commenced against such Person any such proceeding which remains undismissed for a period of sixty (60) days;

(iv) the entrance of any order of relief or other order approving any such case or proceeding involving such Person;

(v) such Person is adjudicated insolvent or bankrupt;

(vi) such Person suffers any appointment of any custodian or the like for it or any substantial part of its property which continues undischarged or unstayed for a period of sixty (60) days;

(vii) such Person makes a general assignment for the benefit of creditors;

(viii) such Person shall fail to pay, or shall state that it is unable to pay, or shall be unable to pay, its debts generally as they become due;

(ix) such Person shall by any act or failure to act, consent to, approve of or acquiesce in any of the foregoing; or

(x) any partnership or corporate action, as the case may be, is taken by such Person for the purpose of effecting any of the foregoing.

Complex clausal structure!

Corporate Loan Agreements

- 2.6 Repayment; Account Control.
- (a) Obligations Due on Maturity. On the Maturity Date, the following Obligations shall become due and payable by Borrower to Lender: (i) the outstanding principal of the Loan, (ii) accrued interest on the Loan, and (iii) any other accrued and unpaid Obligations.

- (b) Account Control. Prior to the Maturity Date, Borrower shall make payments on each Due Date in respect of the Loan, together with all interest, fees, charges and other Obligations owed to Lender, out of all and in an amount equal to all payments, cash, cash equivalents or other Investments or property held by Borrower on such Due Date, together with all interest, fees and charges to which Borrower is entitled in respect thereof, excluding the proceeds of any Excluded Event and any earnings thereon, including all amounts received from Sponsor Member as distributions received by Sponsor Member from Master Holdco or otherwise, and any other proceeds received by Borrower with respect to its Equity Interest in Sponsor Member, including from:

Identifying characteristic dates and associated (universally quantified) logical conditions

Corporate Loan Agreements: Extracting procedures

(i) Loan Value Coverage Ratio. Lender may send written notice to Borrower to request a determination of the Loan Value Coverage Ratio (the "Ratio Determination Request")

(A) at any time from time to time if accompanied by an officer's certificate stating any one or both of the following:

(i) that an event has occurred which has, or could reasonably be expected to have, a Material Adverse Effect or

(ii) that Lender believes in good faith, for reasons specified in reasonable detail in such officer's certificate, that it is reasonable to believe that the Loan Value Coverage Ratio could be lower than 2.25:1 and

(B) any time after ninety (90) days have passed since all Projects have achieved Substantial Completion, for any reason in Lender's sole discretion, but no more frequently than once in each calendar quarter (except in accordance with the previous clause (A)).

If such Loan Value Coverage Ratio is below 2.25:1, Borrower shall within thirty (30) days after such determination make a mandatory prepayment of the Loan in an amount equal to the amount necessary to cause the Loan Value Coverage Ratio, after taking into account such prepayment, to be no lower than 2.25:1.

Within five (5) Business Days after the occurrence of any event that reduces the Percentage Share, Borrower shall deliver to Lender a notice of such event, the cause of the event, and the estimated amount of such reduction in the Percentage Share.

Research Challenges

Rigorous Text Parser Research

- Nature of regulatory text is very different from news:
  - Deontic irrealis mood vs realis mood
  - Far fewer mentions of named entities – and the entities are different (e.g. regulatory agencies and their organs, acts … “the Board”, “employee”, “contractor”, “the Firm”)
  - The meaning of a sentence often relates to text in other parts of the document (e.g. via references, use of definitions).
  - Significant use of abbreviations, references, scoped definitions.
  - Text is much more complex – very low Flesch scores, long complex sentences (many clauses)
  - Text may be structured – single sentence spread across paragraphs, bulleted lists.

- Fortunately, for the most part sentences are dry, precise, declarative and factual or deontic.
  - Text is intended to be clear and descriptive; ambiguity, if present, is deliberate
  - Little, if any, use of metaphors, similes, irony, allusion, sarcasm, satire, alliteration... ➔ we are not dealing with literature (whew!)
- Examples: Programming language reference manuals, math and logic papers, regulatory texts, procedural texts (accounting rules), contracts (employment contracts, loan agreements, ...), prospectuses, RFC for engineering proposals (nuclear power plants)

Cf [Lev 2007] structural semantics
cf [Morgenstern 2014].

Parser research

- Need strategy for dealing with long sentences, spread across multiple paragraphs. Look for techniques to shorten (cf improve Flesch readability metric) e.g.
  - Recognize and mask certain kinds of compound NPs e.g. Tier 2 to Tier 6, Domestic scheduled commercial banks (other than RRBs), branches / Central Processing Centres (CPCs) / Service branches
  - Deal with parenthetical remarks e.g., Tier 1 centres (centres with population of 1,00,000 and above as per 2001 Census)
  - Break up sentences – Sentence expansion.

1) Each depository institution which has a home office or branch office located within a primary metropolitan statistical area, metropolitan statistical area, or consolidated metropolitan statistical area that is not comprised of designated primary metropolitan statistical areas, as defined by the Department of Commerce shall compile and make available, in accordance with regulations of the Bureau, to the public for inspection and copying at the home office, and at least one branch office within each primary metropolitan statistical area, metropolitan statistical area, or consolidated metropolitan statistical area that is not comprised of designated primary metropolitan statistical areas, in which the depository institution has an office the number and total dollar amount of mortgage loans which were (A) originated (or for which the institution received completed applications), or (B) purchased by that institution during each fiscal year (beginning with the last full fiscal year of that institution which immediately preceded the effective date of this title).

#Sentences=1 #words=154 #syllables=329 Flesch score=-126.9

1) We define the designated areas for a depository institution as a primary metropolitan statistical area, metropolitan statistical area, or consolidated metropolitan statistical area that is not comprised of designated primary metropolitan statistical areas, as defined by the Department of Commerce.

Depository institutions with a home office or branch office located in designated areas shall compile and make available per conditions in (1.a) the number and total dollar amount of mortgage loans handled by that institution (as defined in 1.b) during each fiscal year. This obligation begins with the last full fiscal year of that institution which immediately preceded the effective date of this title.
(1.a) The required information shall be compiled and made available to the public for inspection and copying, in accordance with regulations of the Bureau, at the home office and at least one branch office within each designated area.
(1.b) The mortgage loans handled by an institution in a fiscal year are defined as loans which were (A) originated (or for which the institution received completed applications), or (B) purchased by that institution during that year.

#Sentences=9 #words=179 #syllables=349 Flesch score=30.2

Parser research

- Need to customize for our (professional) domains – e.g. handle
  - Different abbreviation styles
  - Quote marks e.g. for "licence" for opening branches (Note: Indian / Japanese English)
  - Conventions for introducing abbreviations e.g. This Evaluation Agreement (“Agreement”)
  - Material in bulleted lists
  - Nested definitions
  - Different citation styles e.g.
    - paragraph 1 (a) (i),
    - Directive 2000/53/EC,
    - referred to in 1) a) i) and ii) above
  - (Inline) Statement and use of definitions e.g.
    - “XXX Confidential Information” means the XXX Samples, specifications, and supporting documentation of the XXX Samples.
    - The term “Confidential Information” shall be used when referring to either party’s or both parties’ Confidential Information, as appropriate.
    - … which were (A) originated (or for which the institution received completed applications), or (B) purchased by that institution …

Parsing Strategy

Article 4: 1. Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of:

(a) all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

(b) portable batteries or accumulators, including those incorporated into appliances, that contain more than 0,002 % of cadmium by weight.

unbullet

Article 4: 1. (a) Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

Article 4: 1. (b) Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of portable batteries or accumulators, including those incorporated into appliances, that contain more than 0,002 % of cadmium by weight.

mask

cleanup

```
prohibit VB ROOT
 +-- Without IN prep
 |   +-- prejudice NN pobj
 |       +-- to IN prep
 |           +-- DirectiveXX NNP pobj
 +-- , , punct
 +-- States NNP nsubj
 |   +-- Member NNP nn
 +-- shall MD aux
 +-- placing NN dobj
    +-- the DT det
    +-- on IN prep ....
```

~~Article 4: 1. (a)~~ Without prejudice to DirectiveX, Member States shall prohibit the placing on the market of all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0.0005 percent of mercury by weight;

~~Article 4: 1. (b)~~ Without prejudice to DirectiveX, Member States shall prohibit the placing on the market of portable batteries or accumulators, including those incorporated into appliances, that contain more than 0.002 percent of cadmium by weight.

Significant amount of pre-processing, custom for each domain

Deep Domain Knowledge Representation: Directed KR

- Regulatory, financial text typically has clear, definite logical content, typically about a specific relation
- Conditions typically refer to predicates defined in a background *domain theory* (i.e. branch/1, population/3, …), and use(numerical) constraints. May be complex (nested disjunctions, conjunctions).
- Directed KR: Questions are unanswerable by mapping into a known vocabulary (cf Rainy Day).
- For wide applicability, system should be able to use textual inference on (formally un-interpreted) text phrases.
- In the worst case, surface appropriate text directly to human. (Support human-in-the-loop mode.)

v) Domestic scheduled commercial banks (other than RRBs) are permitted to open branches, Administrative offices, Central Processing Centres (CPCs) and Service branches in Tier 2 to Tier 6 centres (with population up to 99,999 as per Census 2001 - details of classification of centres tier-wise furnished in Annex 5) and in rural, semi-urban and urban centres in North Eastern States and Sikkim, and to open mobile branches in Tier 3 to Tier 6 centres (with population up to 49,999 as per Census 2001) and in rural, semi-urban and urban centres in North Eastern States and Sikkim without permission from Reserve Bank of India in each case, subject to reporting.

```
rule([rbi2013, 3, v]):
'_ is permitted to open _ in _ with _ '(Bank, Branch, Loc, Condition) :-
  ('domestic scheduled commercial bank'(Bank), \+ 'RRB'(Bank)),
  (branch(Branch);
   'administrative office'(Branch);
   'CPC'(Branch);
   'service branch'(Branch)),
  (((tier(X, Loc), {X >= 2, X=< 6}),
   population(Loc, Pop, 'Census 2001'), {Pop =< 99999});
   (('rural centre'(Loc); 'semi-urban centre'(Loc); 'urban centre'(Loc)),
     state(Loc, State), (State='Sikkim'; 'North Eastern State'(State)))),
  Condition = {'no prior permission needed from RBI', 'subject to reporting'}.
```

<A bank> is permitted to open <a Branch> in <a Location> with <Conditions> if (certain predications hold)

Directed KR

- Regulatory, financial text typically has clear, definite logical content, typically about a specific relation
- KR framework must support ability to name rules (Article 4 1 a), and deny them, and deal with nested exceptions (not illustrated here).

Article 4: 1. Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of:

(a) all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

(b) portable batteries or accumulators, including those incorporated into appliances, that contain more than 0,002 % of cadmium by weight.

2. The prohibition set out in paragraph 1(a) shall not apply to button cells with a mercury content of no more than 2 % by weight.

```
R: '_ prohibits the placing on the market of _'(State, Item):-
        R=rule('Directive 2006/66/EC', ['Article 4', 1, a]),
        'member state'(eu,State),
        'battery or accumulator'(Item),
        applicable(R,Item),
        'mercury content'(Item,'by weight',X percent),
        {X > 0.0005}.
...
R: 'not applicable'(rule('Directive 2006/66/EC', ['Article 4', 1, a]), Item):-
        R=rule('Directive 2006/66/EC', ['Article 4', 2]),
        'button cell'(Item),
        applicable(R,Item),
        'mercury content'(Item, 'by weight', X percent),
        {X =< 2.0}.
```

<A State> prohibits the placing on the market of <an Item> if (certain predications hold)

Directed KR

- Regulatory, financial text typically has clear, definite logical content, typically about a specific relation.
- For wide applicability, system should be able to use textual inference on (formally un-interpreted) text phrases.
- In the worst case, surface appropriate text directly to human. (Support human-in-the-loop mode.)

General Investment Rules
1)a) The Fund may exclusively invest in:
1) a) iii) Recently issued transferable securities and money market instruments, provided that the terms of issue include an undertaking that application will be made for admission to official listing on a Regulated Market and such admission is secured within a year of the issue; and/or

```
'_ may invest in _'(Fund, S) :-
        ('transferable security'(S); 'money market instrument'(S)),
        'issue date'(S, D), reference_date(T), recent(D, T),
        'terms of issue'(S, Terms),
        includes(Terms, 'application will be made for admission to official listing in a Regulated Market'),
        'application for admission to official listing in a Regulated Market'(S, Event),
        date(Event, D), within_a_year(Date, D).
```

<A Fund> may invest in <a Security> if (certain predications hold)

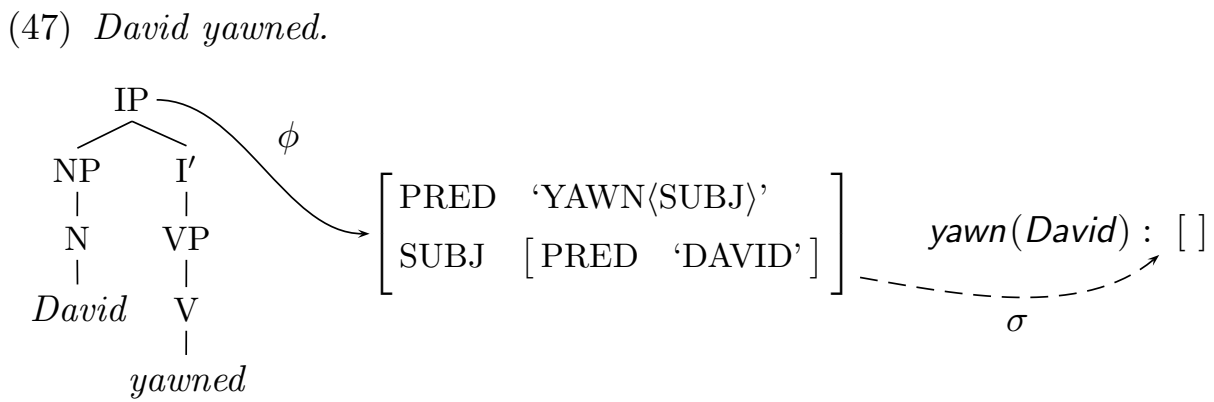
DepGlue

From Dependency Graphs to F-structures

- Starting in the late 70s, extensive work by linguists, logicians, computer scientists has led to deep computational theories of natural languages.
  - Typically, they address complex linguistic phenomena, e.g. long range dependencies.
  - Exemplars: LFG, GPSG, HPSG, CCG, …
- Generally, based on linguistics (non-transformational, cross-language) and logic (type theory, constraints), integrated now with statistical parsers.
- LFG, one of the premier such frameworks, has a well developed framework for semantic analysis (features + glue) that already addresses some of the conceptual difficulties with dependency graphs.
  - CCG is another influential framework
    - focuses on working directly with linear order of words in utterance

Core Proposal: Evolve dependency graphs to F-structures.

Deep Grammar Formalisms: LFG



Constituent structure

Functional structure

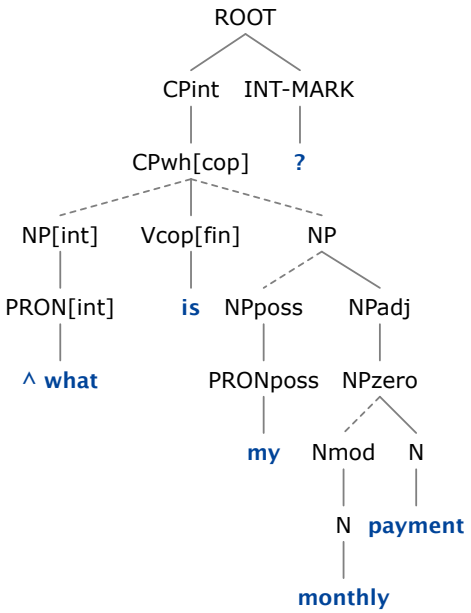
Semantic structure

.

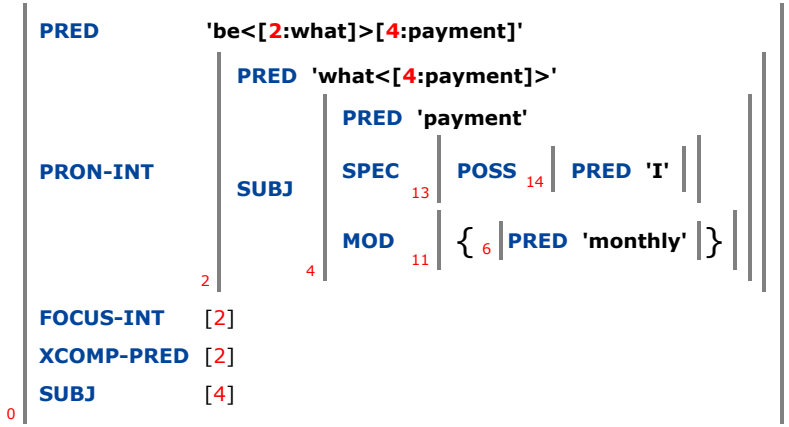
- Deep grammars map strings to meaning representations
  - dependency structures
  - predicate-argument structure
  - simple logical forms
- Lexical-Functional Grammar (LFG): one of the oldest and most well-developed.
  - Organized around lexical and functional structure (not transformational, like Chomsky's work).
  - Simultaneous levels of analysis – structural, functional, semantic, with projections and constraints tying them together

C-structures vs f-structures

C-structure



F-structure



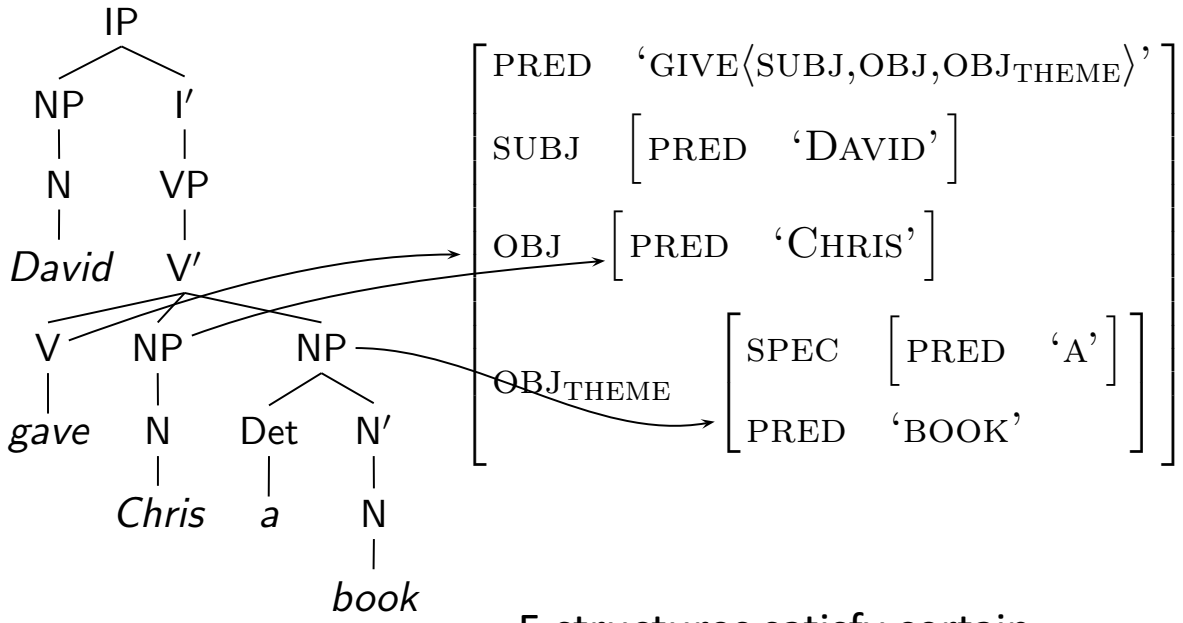
Constituent structure is the overt, more concrete level of linear and hierarchical organization of words into phrases.

- Abstract functional syntactic organization of the sentence, familiar from traditional grammatical descriptions (Subject, Object, Adjunct),
- Representing syntactic predicate-argument structure and functional relations like subject and object.
- Theorized as cross-linguistically uniform.

From [Dalrymple 2009]

F-structures

Complements of Lexical Categories



F-structures satisfy certain semantic constraints: Coherence, Completeness, Consistency.

From [Dalrymple 2009].
See [Dalrymple 2001] for a thorough account

- F-structure reflects the grammatical structure of the sentence (independently of, but coordinated with) c-structure.
- LFG assumes a universally available inventory of grammatical functions:
- SUBJ, OBJ, OBJ_θ, COMP, XCOMP, OBL_θ, ADJ, XADJ
- Θ: semantic roles, such as THEME, SOURCE, GOAL

- F-structures are nested, reentrant attribute value matrices. Values may be sets (with attributes). A very rich vocabulary of descriptions of F-structures has been developed over 30+ years: *equations, disjunctions, negations, optional constraints, (negative) existential constraints , (inside out) functional uncertainty, set descriptions, PCASE (“eval”), non-distributive features*

- Very rich linguistic phenomena modeled using F-structures

From f-structures to logical forms using glue

(74) *David arrived.*

$$f\begin{bmatrix}\text{PRED} & \text{'ARRIVE}\langle\text{SUBJ}\rangle\text{'} \\ \text{SUBJ} & g\big[\text{PRED} \quad \text{'DAVID'}\big]\end{bmatrix}$$

$$[\mathit{arrive}(\mathit{David}), \langle \mathit{David} \rangle] : f_\sigma \otimes \langle g_\sigma \rangle$$

(75) *He yawned.*

$$h\begin{bmatrix}\text{PRED} & \text{'YAWN}\langle\text{SUBJ}\rangle\text{'} \\ \text{SUBJ} & i\big[\text{PRED} \quad \text{'PRO'}\big]\end{bmatrix} \dashrightarrow i_\sigma\big[\text{ANTECEDENT} \quad g_\sigma[\,]\big]$$

$$[\mathit{yawn}(\mathit{David}), \langle \mathit{David}, \mathit{David} \rangle] : h_\sigma \otimes \langle i_\sigma, g_\sigma \rangle$$

Proof:

(81) Meaning constructor premises for *He yawned*:

[context]	$\langle \mathit{David} \rangle$	:	$\langle g_\sigma \rangle$
[yawn]	$\lambda X.\mathit{yawn}(X)$	:	$i_\sigma \multimap h_\sigma$
[he]	$\lambda C.[\mathit{first}(C), \langle \mathit{first}(C), C \rangle]$	:	$\forall C.\langle g_\sigma, C \rangle \multimap [i_\sigma \otimes \langle i_\sigma, g_\sigma, C \rangle]$

(82) **[context-he]** $[\mathit{David}, \langle \mathit{David}, \mathit{David} \rangle] : i_\sigma \otimes \langle i_\sigma, g_\sigma \rangle$

(83) **[context]**, **[he]**, **[yawn]** $\vdash [\mathit{yawn}(\mathit{David}), \langle \mathit{David}, \mathit{David} \rangle] : h_\sigma \otimes \langle i_\sigma, g_\sigma \rangle$

From [Dalrymple 2001]

F-structure provides predicate/arg structure, but not scoped logical forms (w variables, quantifiers etc).

Glue offers a powerful compositional framework for meaning assembly, using deduction in linear logic.

Sematic contributions of components (lambda forms typed with propositional linear logic) are assembled into a term of a given type, via deduction. All terms that can be so constructed represent possible meanings for the utterance.

Glue is agnostic to the actual logic of meanings – one could use Montague's intensional logic, or some other application-dependent logic.

Glue has been shown to be remarkably powerful, handling wide range of semantic phenomena, see [Dalrymple 2001] … , quantification, intensional verbs, modification, coordination, anaphora, ...

Props: From dep parse to meaning representation

- Mask non-core syntactic detail
- Canonicalize and differentiate syntactic constructions
- Mark proposition boundaries
- Propagate relations

Relation
subj - subject *dobj* - direct object *iobj* - indirect object *comp* - complement *prep* - preposition *time* - temporal expressions *prop_of* - adjectival predication *SameAs_arg* - argument of SameAs *outcome* -main clause of conditional *condition* -dependent clause of conditional
mod - modifier *source* - modal-like modification *poss* - possessive
conj - element in conjunction

Meaning representations are directed (possibly cyclic) graphs (represented as JSON with xref/xtarget links).

14 core relations (cf SD ~50)

Getting More Out Of Syntax with PROPS
Gabriel Stanovsky, Jessica Ficler, Ido Dagan, Yoav Goldberg

Props – basic ideas

- Introduce multi-word expressions
- Modality, negation, definiteness, tense, passive/active voice are attribute/value structures
- Appositions → introduce *SameAs_arg* relation
- Existentials → introduce *Exists* node
- Conditional constructions → introduce *condition* and *outcome* relations

- Not dealt with:
  - Quantifiers, scoping

— Text to annotate —

Member States shall prohibit the placing on the market of all batteries or accumulators that contain more than 0.0005 percent of mercury by weight .

— Annotations —

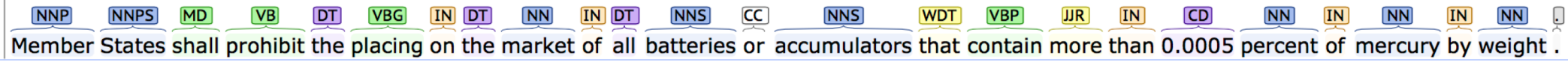
parts-of-speech ✕ named entities ✕ dependency parse ✕ openie ✕

— Language —

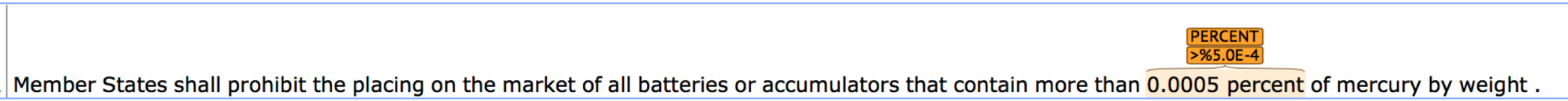
English

Submit

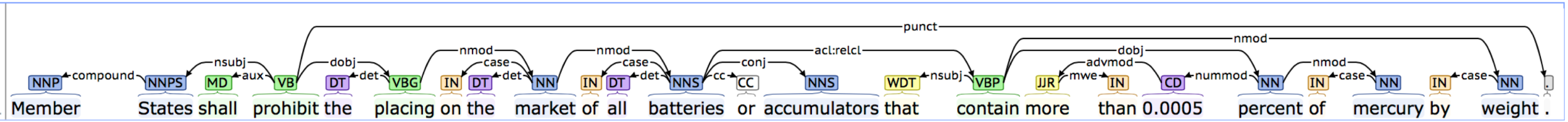
Part-of-Speech:



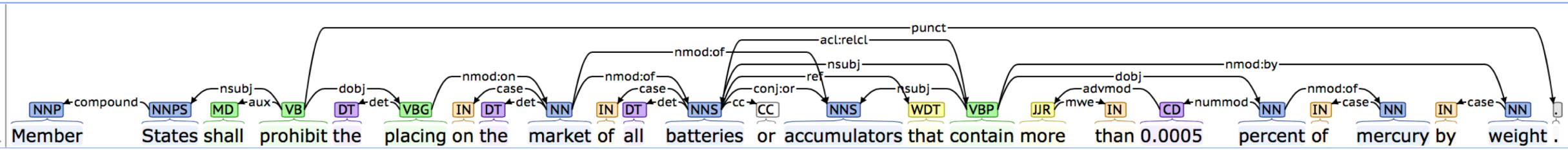
Named Entity Recognition:



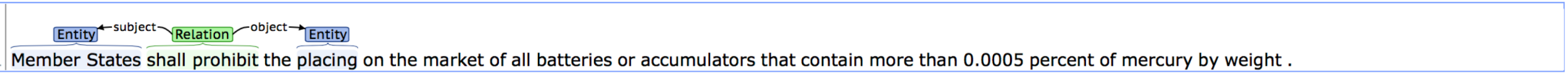
Basic Dependencies:

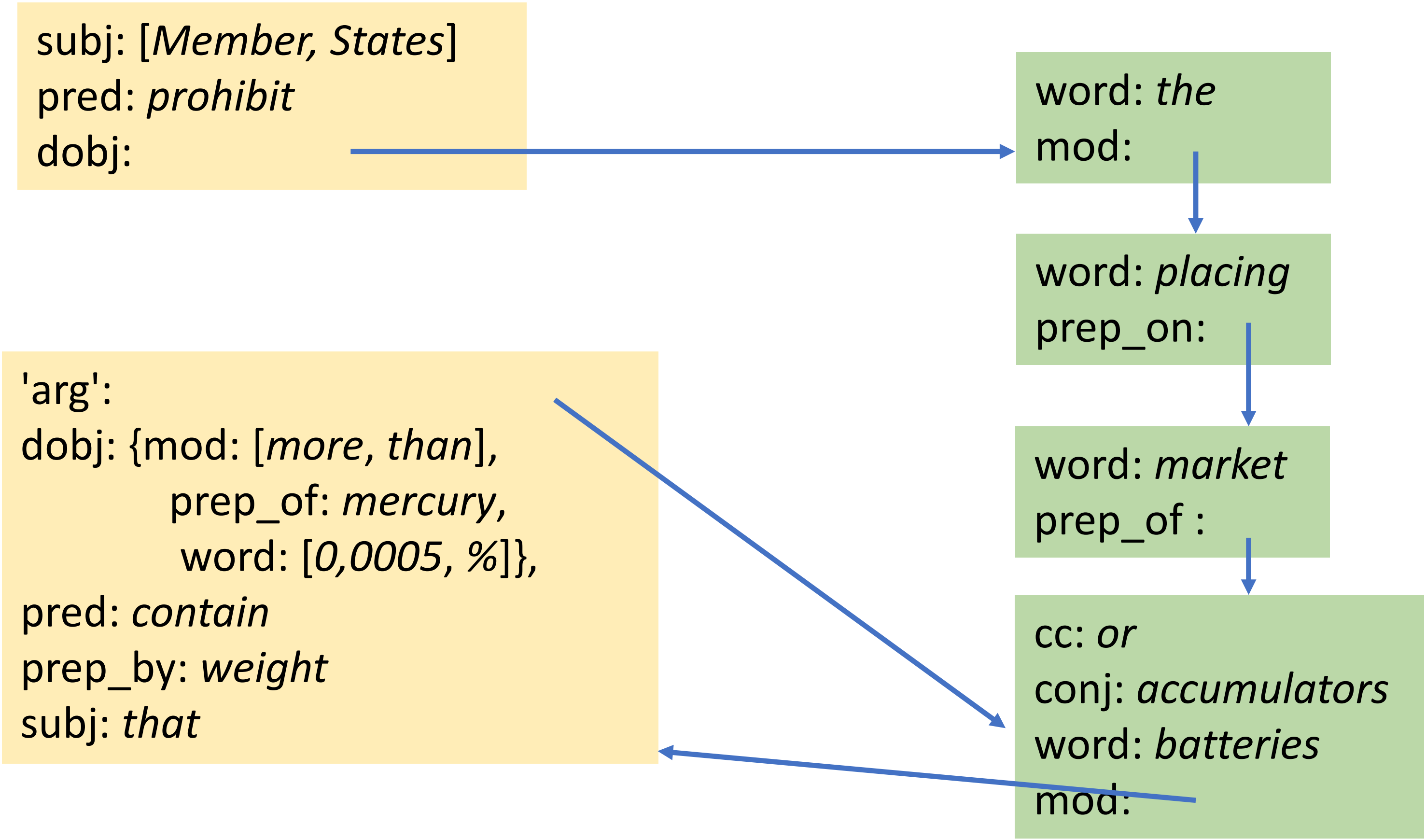


Enhanced++ Dependencies:

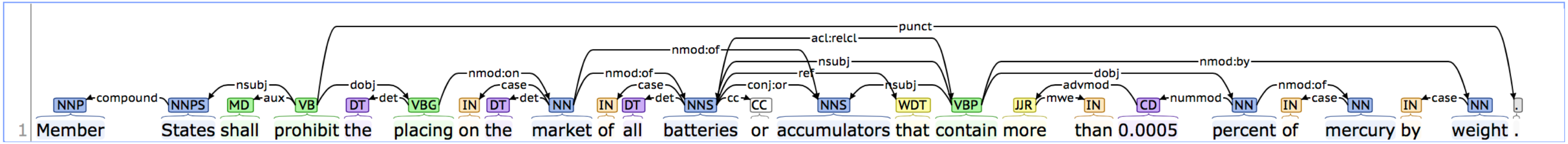


Open IE:





Enhanced++ Dependencies:



Backup

Dependency Parsers

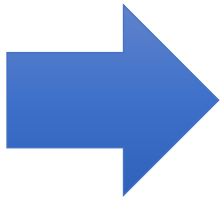
Stanford Dependencies, UD

Google universal POS tags

Google UD

2013, 6 languages; 2014, 11 languages

Interset inerlingua



Universal Dependencies project: Develop Cross-linguistically consistent tree-bank across multiple languages

Provide universal inventory of categories, and guidelines for consistent annotation.

V1.2 has 37 treebanks in 33 languages

Core principle: "Meaning" of a sentence encoded in terms of relation between surface form tokens.

http://universaldependencies.org

What is my monthly payment

```
(l-root w-1-what t-PRON
 (l-cop w-2-is t-VERB)
 (l-nsubj w-5-payment t-NOUN
   (l-nmod:poss w-3-my t-PRON)
   (l-amod w-4-monthly t-ADJ)))
```

What is the monthly rate for this coverage ?

```
(l-root w-1-what t-PRON
 (l-cop w-2-is t-VERB)
 (l-nsubj w-5-rate t-NOUN
   (l-det w-3-the t-DET)
   (l-amod w-4-monthly t-ADJ)
   (l-nmod w-8-coverage t-NOUN
     (l-case w-6-for t-ADP)
     (l-det w-7-this t-DET)))
 (l-punct w-9-? t-PUNCT))
```

UDepLambda [Reddy 2016]

```
(l-root w-1-what t-PRON
 (l-cop w-2-is t-VERB)
 (l-nsubj w-5-rate t-NOUN
   (l-det w-3-the t-DET)
   (l-amod w-4-monthly t-ADJ)
   (l-nmod w-8-coverage t-NOUN
     (l-case w-6-for t-ADP)
     (l-det w-7-this t-DET)))
 (l-punct w-9-? t-PUNCT))
```

What is the monthly rate for this coverage ?

```
(l-punct
 (l-nsubj
  (l-cop w-1-what w-2-is)
  (l-nmod
   (l-det
    (l-amod w-5-rate w-4-monthly) w-3-the)
   (l-case (l-det w-8-coverage w-7-this) w-6-for)))
 w-9-?)
```

Binarize

Obliqueness hierarchy:
punct < nsubj < cop

```
(lambda $0:<a,e>  (exists:ex  $1:<a,e>  (exists  $2:<a,e>
   (and:c
    (p_TYPE_w-1-what:u $0)
    (p_EVENT_w-1-what:u $0)
    (p_EVENT.ENTITY_arg0:b $0 $0)
    (p_TARGET:u $0)
    (p_TYPE_w-5-rate:u $1) (p_EVENT_w-5-rate:u $1)
    (p_EVENT.ENTITY_arg0:b $1 $1)
    (p_TYPEMOD_w-4-monthly:u $1)
    (p_EMPTY:u $1)
    (p_TYPE_w-8-coverage:u $2) (p_EVENT_w-8-coverage:u $2)
    (p_EVENT.ENTITY_arg0:b $2 $2)
    (p_EMPTY:u $2)
    (p_EVENT.ENTITY_l-nmod.w-6-for:b $1 $2)
    (p_EVENT.ENTITY_arg1:b $0 $1)))))
```

Subsitute, Simplify

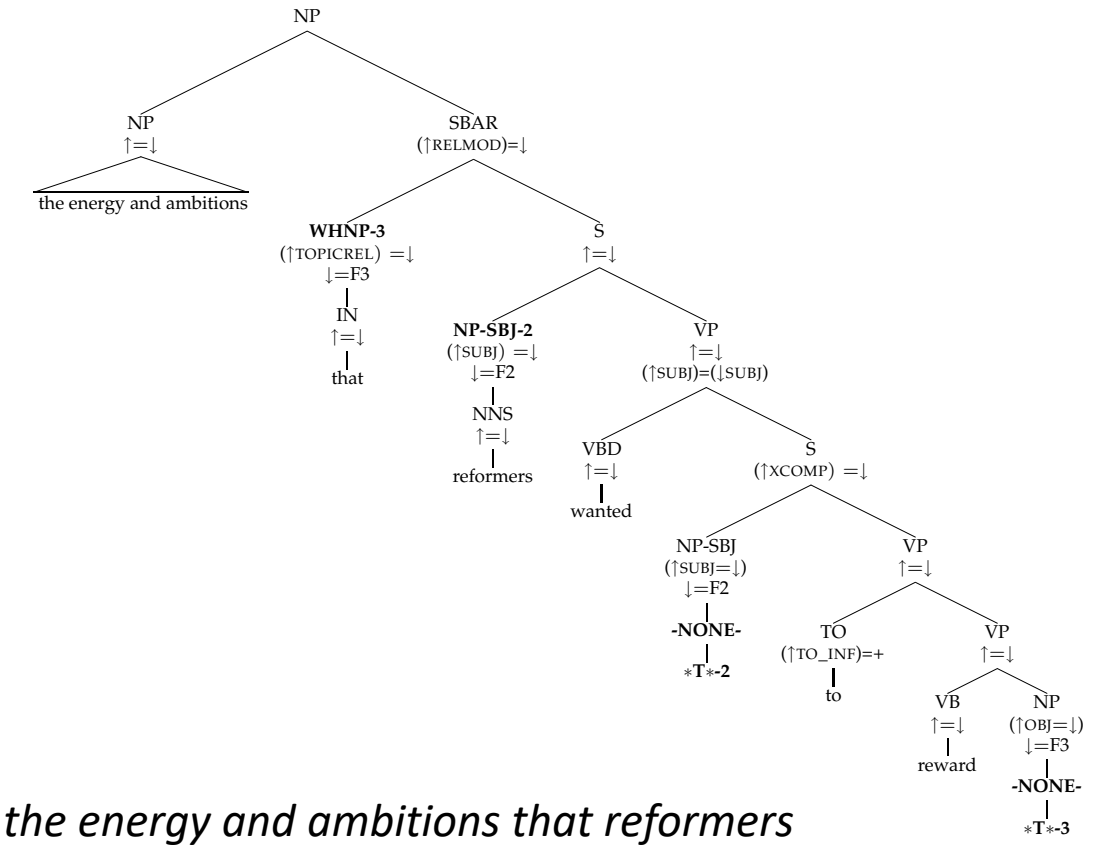
QUESTION(what:x), what.arg0(what:x), what.arg1(rate:x), tmod(monthly,rate:x), coverage.arg0(coverage:x), rate.nmod.for(coverage:x),

~~rate.arg0(rate:x)~~

Transform to application-friendly syntax

Generated using Stanford CoreNLP 3.7

What dependencies do not capture I/III



the energy and ambitions that reformers wanted to reward

root(ROOT-0, energy-2)
det(energy-2, the-1)
cc(energy-2, and-3)
conj:and(energy-2, ambition-4)
mark(wanted-7, that-5)
nsubj(wanted-7, reformers-6)
nsubj:xsubj(reward-9, reformers-6) // *T*-2
dep(energy-2, wanted-7)
mark(reward-9, to-8)
xcomp(wanted-7, reward-9)

Missing:
dobj:xsubj(reward-9, that-5) // *T*-3

Long Distance Dependencies

What dependencies do not capture II/III

```
(l-root w-2-wants t-VERB
 (l-nsubj w-1-everybody t-NOUN)
 (l-xcomp w-4-buy t-VERB
   (l-mark w-3-to t-PART)
   (l-dobj w-6-house t-NOUN
     (l-det w-5-a t-DET)))
 (l-punct w-7-. t-PUNCT))
```

Everybody wants to buy a house.

Adding **(l-xcomp w-4-buy t-VERB (l-nsubj w-1-everybody t-NOUN))**
actually gives the meaning of *Everybody wants that everybody buys a house*

```
all(X, person(X),
    wants(X, a(Y,
      house(Y)&
          buys(X,Y))))
```

[Schuster & Manning 2016]

```
(l-root w-2-sleeps t-VERB
 (l-nsubj w-1-everybody t-NOUN)
 (l-cc w-3-or t-CONJ)
 (l-conj w-5-awake t-ADJ
   (l-cop w-4-is t-VERB))
 (l-punct w-6-. t-PUNCT))
```

Everybody sleeps or is awake.

Adding **(l-conj w-5-awake t-VERB (l-nsubj w-1-everybody t-NOUN))**
actually gives the meaning of *Everybody sleeps or everybody is awake*

```
all(X, person(X),
    sleeps(X) | awake(X)))
```

Need to introduce variables (or some other linguistic device), i.e. move beyond core principle of dependency parsing.

Generalized Quantifiers

Quick aside: Quantifiers arise in regs!

Member States shall prohibit the placing on the market of all batteries or accumulators, that contain more than 0.005 percent of mercury by weight.

```
(l-root w-4-prohibit t-VERB
 (l-nsubj w-2-states t-PROPN
  (l-compound w-1-member t-PROPN))
 (l-aux w-3-shall t-AUX)
 (l-dobj w-6-placing t-VERB
  (l-det w-5-the t-DET)
  (l-nmod w-9-market t-NOUN
   (l-case w-7-on t-ADP)
   (l-det w-8-the t-DET)
   (l-nmod w-12-batteries t-NOUN
    (l-case w-10-of t-ADP)
    (l-det w-11-all t-DET)
    (l-cc w-13-or t-CONJ)
    (l-conj w-14-accumulators, t-NOUN)
    (l-acl:relcl w-16-contain t-VERB
     (l-nsubj w-15-that t-PRON)
     (l-dobj w-20-percent t-NOUN
      (l-nummod w-19-0.005 t-NUM
       (l-advmod w-17-more t-ADJ
        (l-mwe w-18-than t-ADP)))
      (l-nmod w-22-mercury t-NOUN
       (l-case w-21-of t-ADP)))
     (l-nmod w-24-weight. t-NOUN
      (l-case w-23-by t-ADP)))))))
```

```
R: '_ prohibits the placing on the market of _'(State, Item):-
        R=rule('Directive 2006/66/EC', ['Article 4', 1, a]),
        'member state'(eu,State),
        'battery or accumulator'(Item),
        applicable(R,Item),
        'mercury content'(Item,'by weight',X percent),
        {X > 0.0005}.
```

Generalized Quantifiers

What dependencies do not capture III/III

```
(l-root w-5-carrying t-VERB
 (l-nsubj w-1-sue t-PROPN
   (l-cc w-2-and t-CONJ)
   (l-conj w-3-mary t-PROPN))
 (l-aux w-4-are t-AUX)
 (l-dobj w-7-piano t-NOUN
   (l-det w-6-a t-DET))
 (l-punct w-8-. t-PUNCT))
```

```
a(X,  piano(X),
      carrying({sue,mary}, X))
```

```
a(X,  piano(X), carrying(sue, X)) &
a(X,  piano(X), carrying(mary, X))
```

Sue and Mary are carrying a piano.

Adding **(l-root w-5-carrying t-VERB (l-nsubj w-1-sue t-PROPN))**
actually gives the distributive interpretation:
Sue is carrying a piano and Mary is carrying a piano.

Need to support non-distributive interpretations of conjoined subjects

[Schuster & Manning 2016]

Conjoined subjects

Core Linguistics to Knowledge projects

Parser project:
Get better dependency parsers customized for our domains.
Extend tree-banks.
Implement f-description annotations for tree banks

(Interactive) Sentence expansion → raise Flesch score (cf compression)

(Interactive) Sentence simplification (cf compression) – dropping expletives, elaborations

Semantic forms to knowledge representation (cf TIL2)

(Interactive) Sentence comprehension – surface ambiguities to user for action.

Corpus generation for supervised techniques.

Bag of phrases model – nearest neighbor metric.

Shallower OpenIE techniques for Knowledge construction (Need to be adapted for professional domains.)

Deep Obligation extraction, comparison, mapping.

Note: need to support experienced users, and novice users.

Knowledge Representation and Reasoning Infrastructure

Continuous Ingestion of documents

Scaled out representation of generated assertions, indexing, meta-data (e.g. provenance, dependency support).

Knowledge Extraction

Directed Reasoning, support for domain theories

Compositional (contextual) reasoning

Textual Reasoning

Explanation generation (w/ built-in support for correction)