Towards multilingual text generation from structured formal representations

Dana Dannélls

NLP Research Unit
Department of Swedish Language
University of Gothenburg
Graduate School of Language Technology (GSLT)
dana.dannells@svenska.gu.se

October 21, 2009

Natural Language Generation

What is Natural Language Generation (NLG)

- the process of mapping internal representations of information into human languages whether textual or spoken
  - Input: non-linguistic source of information
  - Output: written documents, summaries, reports, etc.
- aims to develop a computational theory of human language ability

Multilingual language generation (MLG)

- produce texts automatically in various languages
- an alternative approach to machine translation

Key problems of NLG

A problem of decision making under multiple constraints:

- propositional knowledge at hand
- linguistic tools available
- communicative goals and intentions to be achieved
- audience the text is aimed at
- present and past discourse

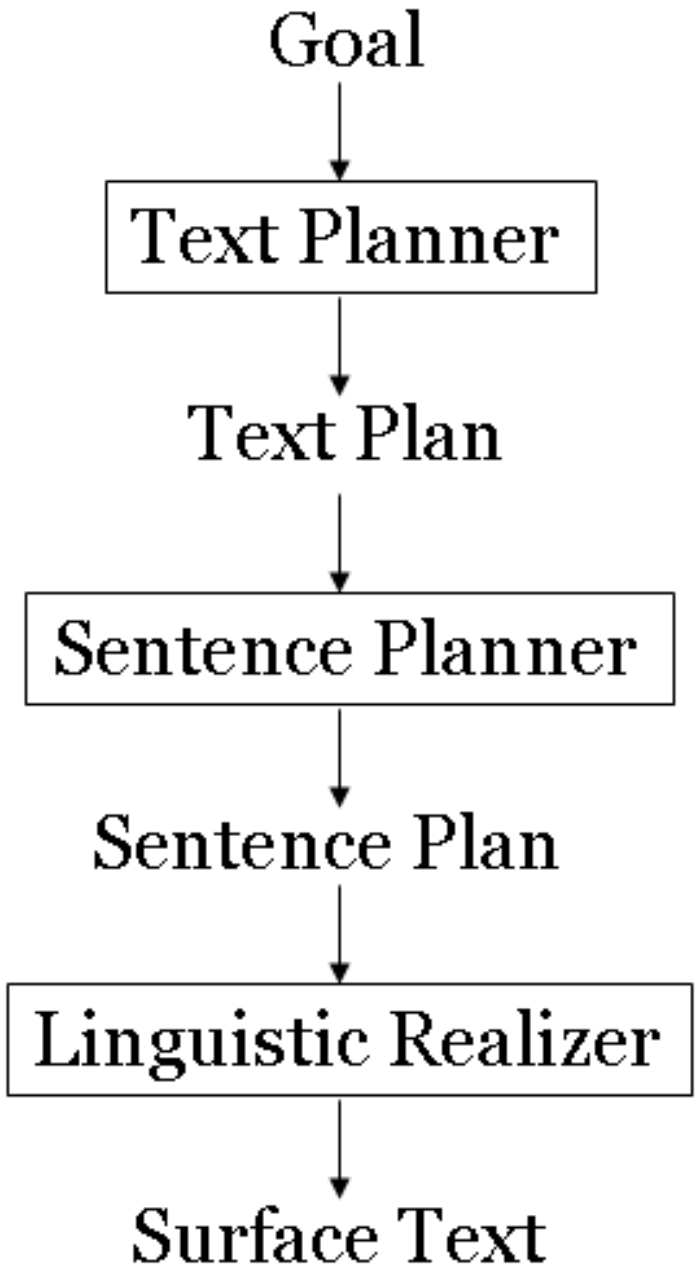
NLG system architecture

- Text planner
  - Content determination
  - Discourse planning
- Sentence planner
  - Sentence aggregation
  - Lexicalization
  - Referring expression generation
- Linguistic realisation

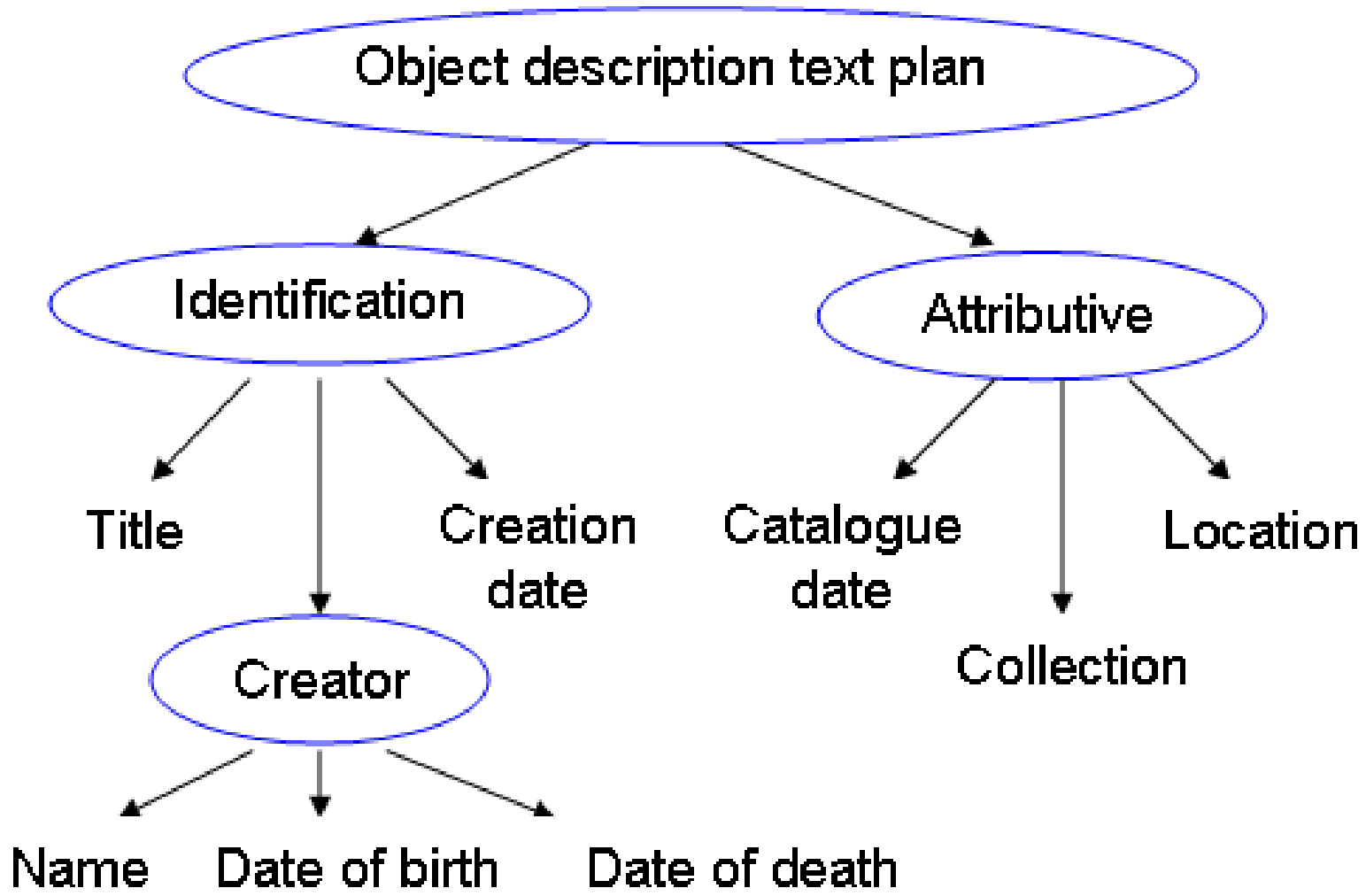
Computational approaches

- Text planning
  - Rhetorical Structure Theory (Mann and Thompson, 1988)
  - Schema-based (McKeown, 1985)
  - Machine learning and statistical (Barzilay et al., 2002)
- Sentence planning
  - Semi-automatic (Shaw, 1998; Dalianis, 1996)
  - Statistical (Zhong and Stent, 2003; Belz et al., 2007)
  - Corpus based (Karamanis and Mellish, 2005)
- Linguistic realisation
  - Template based (Busemann and Horacek, 1998; McRoy et al., 2003)
  - Grammar based (Goldberg et al., 1994; Bateman et al., 1998)
  - Hybrid (Williams et al., 2007)

NLG system architecture (Reiter and Dale, 2000)



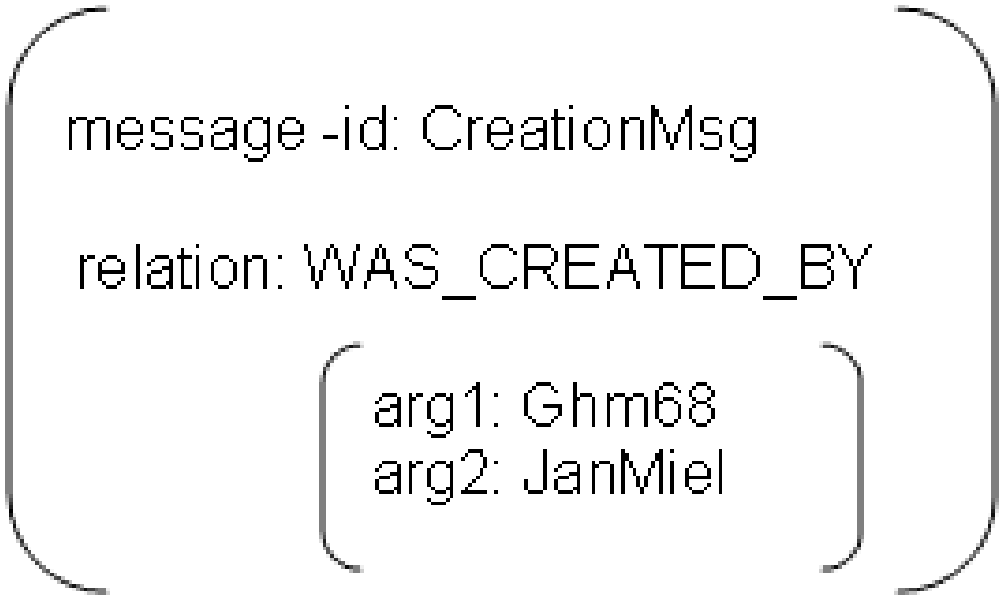
Text plan example



Rest on the Hunt was painted by Jan Miel (1599-1664) in 1642. The Ghm68, catalogued 1660, belongs to Martin Heade's collection that is located in Amon Carter Museum, Texas.

Message example

A message forms an abstraction that mediates between the data structures used in the underlying application and the eventual texts to be generated.



A *was created by* message.

Sentence plan example

type: SentenceAbstractSpecification

object: Ghm68
relation1: has_title
value: |Rest on the Hunt|

relation2: was_created_by
value: |Jan Miel|

relation3: was_born_died_in
valueBirth: |1599|
valueDeath: |1664|

relation4: has_time_span
value: |1642|

Rest on the Hunt was painted by Jan Miel (1599-1664) in 1642.

Aggregation

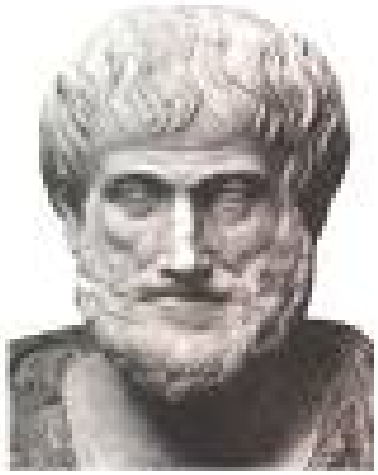
In linguistics
aggregation is *the compositional building of a linguistic structure of a sentence.*

Coordination, extraction, ellipsis, focus ,centering, discourse and lexical semantics

In computational linguistics
aggregation is *an organizing principle that takes place on different levels.*

Coordination, subordination, omisson, substation, combination, organization

The traditional notion of Ontology



- a branch of metaphysics which aims to discover what entities exist and attempts to sort these entities into categories
- built from the Greek forms onto- 'of being' and logos 'speech, reason'

Ontologies in computer science

> *An ontology is an explicit specification of concepts and the relations among them. It provides a shared understanding of some domain of interest and a formal vocabulary for information exchange. (Gruber, 1993)*

Some types of ontologies

- Top level ontologies
- Domain ontologies
- Linguistic ontologies

The richness of their internal structure

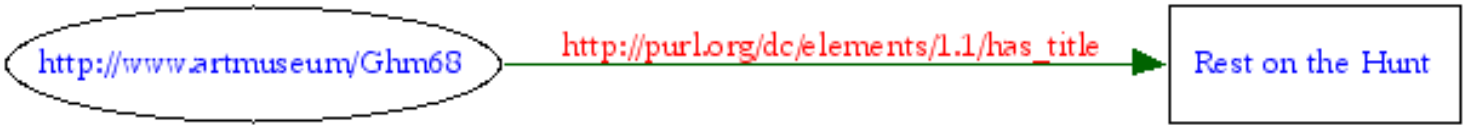
- Lightweight: concepts, relationships and properties
- Heavyweight: lightweight + axioms and constraints

The Semantic Web (Berners-Lee et al., 2001)

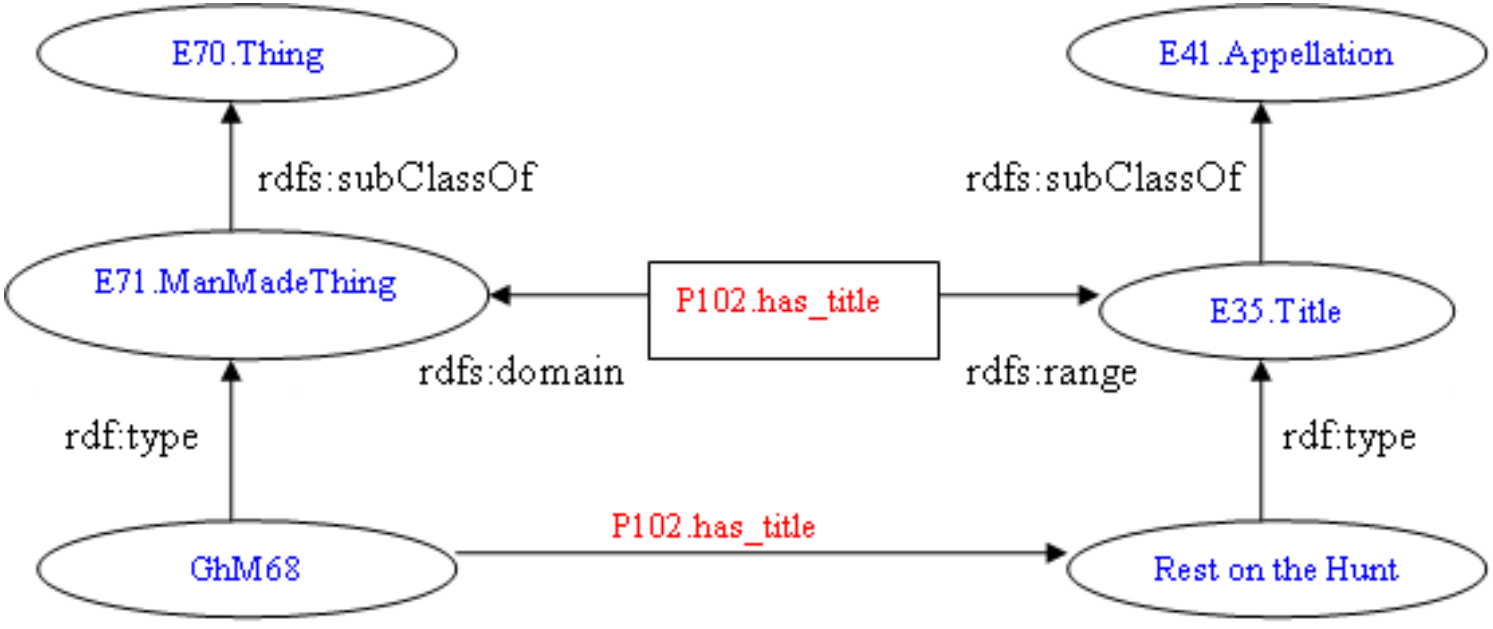
- give information explicit meaning that machines can understand
- support a distributed Web at the level of the data rather than at the level of the presentation
- using global references called Uniform Resource Identifiers (URI)

Semantic Web modeling languages

- The Resource Description Framework (RDF) provides meta-data about Web resources



- RDF Schema (RDFS) defines the primitives for creating ontologies



The Web Ontology Language (OWL)

- contains a richer set of operators
  - inverseOf, sameAs
  - unionOf, complementOf, intersectionOf
- has computational properties for reasoning systems
- different levels of expressivity
  - OWL Lite
  - OWL DL (Description Logic)
  - OWL Full

Semantic Web ontologies and NLG

- provide a wide variety of modeling capabilities for relating and presenting different kinds of information
- offer an unlimited amount of high level semantic specification
- information access to distributed repositories
- provide a basis for the underlying textual content

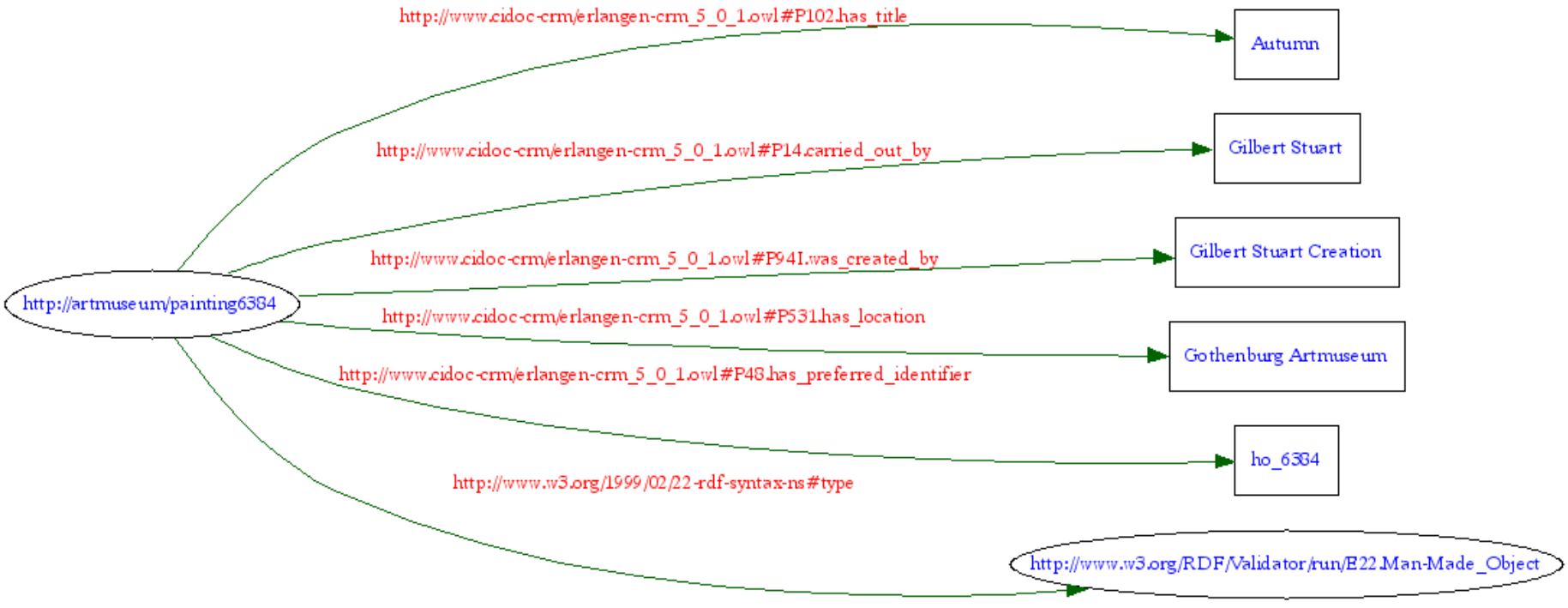
  crm:P97293 rdf:subject cidoc-crm:Householder ;
  rdf:predicate cidoc-crm:represents ;
  rdf:object cidoc-crm: MariaBanck.

- expose semantic web services directly in response to natural language queries, i.e., SPARQL

NLG from Semantic Web ontologies

- RDF/XML (Wilcock, 2003)
  - spoken summaries and explanations
- DAML+OIL (Wilcock and Jokinen, 2003)
  - dialogue responses
- RDF/OWL – MIAKT (Bontcheva and Wilks, 2004)
  - textual summaries
- RDF/OWL – ONTOSUM (Bontcheva, 2005)
  - textual summaries, oriented towards the user
- OWL – ILEX (O'Donnell et al., 2001)
  - personalised object descriptions
- OWL-DL (Mellish and Pan, 2008)
  - focus on content selection

An example - NLG from SW ontologies



Painting_6384 has title Autumn. Painting_6384 was created by Gilbert Stuart. Painting_6384 has location Gothenburg Artmuseum.

This painting is titled Autumn, **it** was painted by Gilbert Stuart **and it** is located in Gothenburg Artmuseum.

The research problems

Currently available NLG components suffer from one or more of the following problems:

- sentence planning is usually embedded in a system and cannot be re-used by other applications;
- existing approaches to sentence planning do not take advantage of the full expressive power of language;
- systems employ direct verbalisation that is close to the ontology structure;
- aggregation has been employed in relatively simple forms using superficial approaches;
- systems that generate from Semantic Web ontologies are monolingual.

The research goals

The primary goal
To develop a sentence planner that is able to make linguistically motivated decisions in manipulating non-linguistic abstract representations.

The secondary goal
To develop a module that uses Semantic Web ontology representations and is capable of utilizing some of the expressive power of Semantic Web languages.

Our guiding hypothesis is that aggregation possibilities are different in different languages.

Resources required for generation

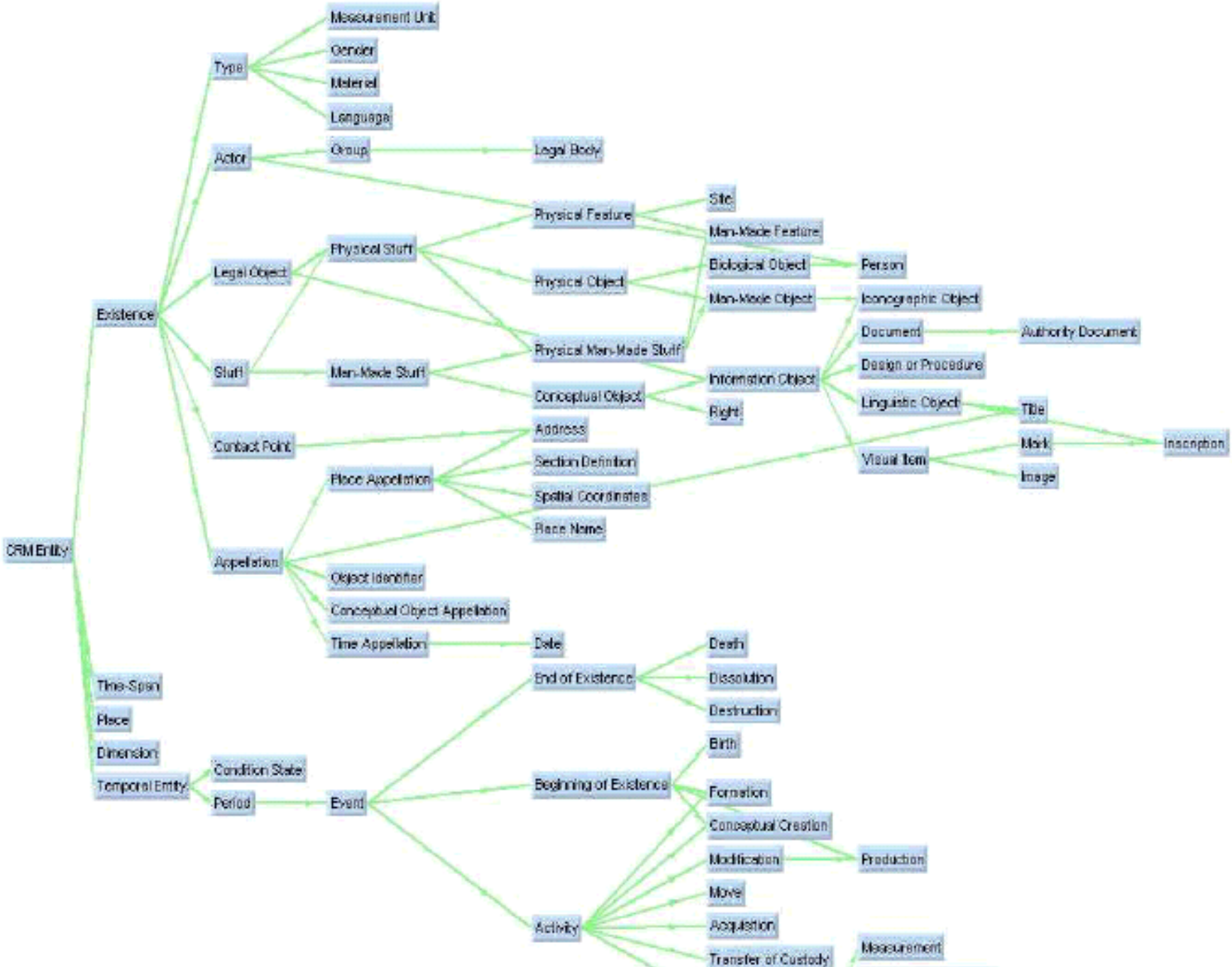
- Domain ontology
- Discourse plan
- Lexicon
- Grammar

The domain ontology

The CIDOC Conceptual Reference Model (CIDOC CRM), developed by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM) (Crofts et al., 2008).

- ISO standard since 2006
- Based on data structures in museum applications
- It contains 87 classes and 260 relationships
- Available in OWL

The CIDOC-CRM



Discourse plan (McKeown, 1985)

- Impose a discourse structure upon occurring sets of predicates
- A rhetorical schema describes the discourse structure
- Rhetorical predicates characterize the predicting acts a speaker may use for making different statements

Description schema:

Describe–Object $->$
Identification–Property,
Attributive–Property.

Identification–Property $->$
$\{T1 / \{T2 / T3\}\}$.

Attributive–Property $->$
$\{T4 / T5\}$.

Discourse plan

Template specifications:

[T1] (a) object's title | (b) object's creator | (c) creation date
[T2] (a) creator date of birth | (b) creator date of death
[T3] (a) object id | (b) object material | (c) object size
[T4] (a) owner | (b) location | (c) catalogue date | (d) collection
[T5] (a) object's identifier | (b) identified place

Example:

[*T1b*] *Thomas Sully painted this half-length* [*T1a*] *Portrait of Queen Victoria* [*T1c*] *in 1838. The subject is now installed in the* [*T4d*] *Wallace Collection,* [*T4b*] *London.*

Electronic lexical resources

- Swedish
  The PAROLE and SIMPLE lexicons (Lenci et al., 2000)
- English
  WordNet and FrameNet (Fillmore et al, 2003)
- Hebrew
  WordNet (Ordan and Wintner, 2005)

The Grammatical Framework (GF)

A functional grammar formalism, oriented towards multilingual grammar development and generation (Ranta, 2004).

Differentiates between domain dependent and domain independent linguistic resources

- Abstract syntax is a set of functions and categories (Martin-Löf, 1984)
- Concrete syntax defines the linearizations of functions and categories into strings
- Resource Grammar Library provides an API for 12 languages

Data

Written object descriptions in Swedish, English and Hebrew

- National galleries catalogues
- Painting and portrait collections
- E-libraries and open archives

OWL representations of cultural objects

- Manually created by the author
- Each file contains about 100 RDF statements

Text linguistic analysis process

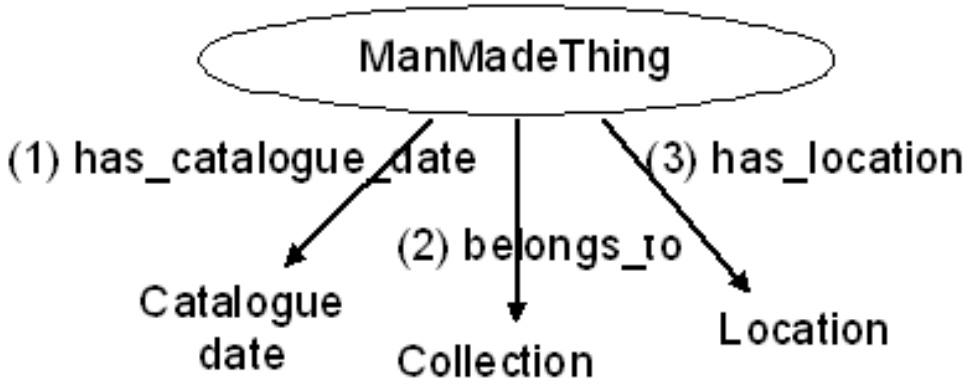
Given a set of statements and a set of written object descriptions in English, Swedish and Hebrew

For each language

- recognize given statements in the text
- examine how they are related and realized in the discourse
    - semantic constituents
    - syntactic constructions
    - reference items
- capture language dependent aggregation rules

Gradually reduce language dependent rules by writing language independent ones

Language dependent aggregation



(1) The subject made its first appearance in 1880, (3) now installed in the East Gallery of the American Wing Courtyard. (2) It belongs to Morris K. Jesup's collection.

(1) Först på 1900 talet kom den till Sverige och (3) hänger nu på Gripsholms slott (2) i Statens porträttsamling.

(1) ha-tmuwnah hegieh larisunah le-Aeretz yisraAel besnat 1960. (2) hyA sayeket le-quwleqitzyah sel Amir bachar (3) se-nimtzet be-muwzeyAuwn haAretz be-tel Aabiyb

Non–linguistic abstract representations

Rest on the Hunt was painted by John Miel in 1642.
Rastande jägare är målad av John Miel år 1642.
menuhat tzayyd tzuwyrah 'al yedey guwn miyAel be-1642.

Abstract syntax

cat
Identification ; ConceptA ; CreationProperty ; ConceptB ; CreationV ;
IndividualFunction ; Individual ;
fun
ObjDescription: ConceptA → CreationProperty → ConceptB →
Identification ;
CreationAct: CreationV → ConceptB → CreationProperty ;
SubjectFunction: IndividualFunction → ConceptA → ConceptB ;
SubjectName: Individual → ConceptA ;
RestOnTheHunt, JanMiel, 1642 : Individual ;
WasPainted : CreationV ;
ByPerson, InYear : IndividualFunction ;

Proposed experiments

The generated object descriptions

- compared to original object descriptions written by experts
- evaluated by native speakers of the language

The sentence planner module

- applied to other ontologies from other domains
- tested with other languages

State of the work

- We have started the development of a computational sentence planner module
- The module addresses both internal composition of clauses and their organization into larger structure
- Text linguistic approach is taken in order to study differences across languages
- We already identified some semantic and syntactic differences in how statements are aggregated

Expected contributions

- A sentence planner module that can be re-used by other applications
- An approach that addresses both clause organisations into larger structures and their internal compositions
- Grammar driven generation that utilizes a lexicon
- Linguistically motivated aggregations
- Multilingual prototype that is able to generate from Semantic Web ontologies

Thank You

Thank you for listening

References

- Bateman, J. and Sharoff, S. 1998. *Multilingual grammars and multilingual lexicons for multilingual text generation.* In Multilinguality in the lexicon II, ECAI.
- Bontcheva, K. 2005. *Generating tailored textual summaries from ontologies*. In Second European Semantic Web Conference (ESWC), pages 531–545.
- Bontcheva, K. and Wilks, Y. 2004. *Automatic report generation from ontologies: the MIAKT approach.* In Proceedings of the Nineth International Conference on Applications of Natural Language to Information Systems (NLDB), pages 324–335.
- Hercules, D. 1996. *Concise Natural Language Generation from Formal Specifications.* PhD thesis, Stockholms University.
- Fillmore, C. J., Johnson, C. R., and Petruck, M. R. 2003. Background to framenet. *International Journal of Lexicography*, 16(3):235–250.
- Lenci et al. 2000. Simple: A general framework for the development of multilingual lexicons. *Lexicography,* 13(4):249–263.
- McKeown, K. R. 1985. *Text generation : using discourse strategies and focus constraints to generate natural language text*. Cambridge University Press.
- O'Donnell, M. J., Mellish, C., Oberlander, J., and Knott, A. 2001. ILEX: An architecture for a dynamic hypertext generation system. *Natural Language Engineering,* 7(3): 225–250.
- Ranta, A. 2004. Grammatical framework, a type-theoretical grammar formalism. *Journal of Functional Programming,* 14(2):145–189.
- Reape, M. and Mellish, C. 1999. *Just what is aggregation anyway?* In Proceedings of the 7th European Workshop on Natural Language Generation, pages 20–29.
- Reiter, E. and Dale, R. 2000. *Building Natural Language Generation Systems.* MIT Press and The McGraw-Hill Companies, Inc.
- Shaw, J. 1998. *Clause aggregation using linguistic knowledge.* In Proceedings of the 9th INLG.
- W3C. Owl web ontology language overview. 2004. URL: http://www.w3.org/TR/owl-features/.
- Wilcock, G. 2003. *Talking owls: Towards an ontology verbalizer*. In Human Language Technology for the Semantic Web and Web Services, pages 109–112.
- Wilcock, G. and Jokinen, K. 2003. *Generating responses and explanations from RDF/XML and DAML+OIL.* In Knowledge and Reasoning in Practical Dialogue Systems IJCAI, pages 58–63.