On generating coherent multilingual descriptions of museum objects from Semantic Web ontologies

Dana Dannélls
Språkbanken
Department of Swedish
University of Gothenburg, Sweden
dana.dannells@svenska.gu.se

Abstract

During the last decade, there has been a shift from developing natural language generation systems to developing generic systems that are capable of producing natural language descriptions directly from Web ontologies. To make these descriptions coherent and accessible in different languages, a methodology is needed for identifying the general principles that would determine the distribution of referential forms. Previous work has proved through cross-linguistic investigations that strategies for building coreference are language dependent. However, to our knowledge, there is no language generation methodology that makes a distinction between languages about the generation of referential chains. To determine the principles governing referential chains, we gathered data from three languages: English, Swedish and Hebrew, and studied how coreference is expressed in a discourse. As a result of the study, a set of language specific coreference strategies were identified. Using these strategies, an ontology-based multilingual grammar for generating written natural language descriptions about paintings was implemented in the Grammatical Framework. A preliminary evaluation of our method shows language-dependent coreference strategies lead to better generation results.

createdBy (Guernica, PabloPicasso)
currentLocation (Guernica, MuseoReinaSofía)
hasColor (Guernica, White)
hasColor (Guernica, Gray)
hasColor (Guernica, Black)

Guernica is created by Pablo Picasso. **Guernica** has as current location the Museo Reina Sofía. **Guernica** has as color White, Gray and Black.

Figure 1: A natural language description generated from a set of ontology statements.

1 Introduction

During the last decade, there has been a shift from developing natural language generation systems to developing generic systems that are capable of producing natural language descriptions directly from Web ontologies (Schwitter and Tilbrook, 2004; Fuchs et al., 2008; Williams et al., 2011). These systems employ controlled language mechanisms and Natural Language Generation (NLG) technologies such as discourse structures and simple aggregation methods to verbalise Web ontology statements, as exemplified in figure 1.

If we want to adapt such systems to the generation of coherent multilingual object descriptions, at least three language dependent problems must be faced, viz. lexicalisation, aggregation and generation of referring expressions. The ontology itself may contain the lexical in-

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 76–84,
Utica, May 2012. ©2012 Association for Computational Linguistics

Guernica is created by Pablo Picasso.
It has as current location the Museo Reina Sofía.
It has as color White, Gray and Black.

Guernica målades av Pablo Picasso.
Den finns på Museo Reina Sofía.
Den är målad i vitt, svart och grått.

Figure 2: A museum object description generated in English and Swedish.

formation needed to generate natural language (McCrae et al., 2012) but it may not carry any information either about the aggregation of semantic concepts or the generation of a coherent discourse from referring expressions. Halliday and Hasan (1976), and other well known theories such as Centering Theory (Grosz et al., 1995), propose establishing a coherent description by replacing the entity referring to the Main Subject Reference (MSR) with a pronoun – a replacement which might result in simple descriptions such as illustrated in figure 2. Although these descriptions are coherent, i.e. they have a connectedness that contributes to the reader's understanding of the text, they are considered non-idiomatic and undeveloped by many readers because of consecutive pronouns – a usage which in this particular context is unacceptable.

Since previous theories do not specify the types of linguistic expressions different entities may bear in different languages or domains, there remain many open questions that need to be addressed. The question addressed here is the choice of referential forms to replace a sequence of pronouns, which makes the discourse coherent in different languages. Our claim is that different languages use different linguistic expressions when referring to a discourse entity depending on the semantic context. Hence a natural language generator must employ language dependent coreferential strategies to produce coherent descriptions. This claim is based on cross-linguistic investigations into how coreference is expressed, depending on the target language and the domain (Givón, 1983; Hein, 1989; Ariel, 1990; Prince, 1992; Vallduví and Engdahl, 1996).

In this paper we present a contrasting study conducted in English, Swedish and Hebrew to learn how coreference is expressed. The study was carried out in the domain of art, more specifically focusing on naturally-occurring museum object descriptions. As a result of the study, strategies for generating coreference in three languages are suggested. We show how these strategies are captured in a grammar developed in the Grammatical Framework (GF).[1] We evaluated our method by experimenting with lexicalised semantic web ontology statements which were structured according to particular organizing principles. The result of the evaluation shows language-dependent coreference strategies lead to better generation results.

2 Related work

Also Prasad (2003) employed a corpus-based methodology to study the usage of referring expressions. Based on the results of the analysis, he developed an algorithm to generate referential chains in Hindi. Other algorithms for characterizing referential expressions based on corpus studies have been proposed and implemented in Japanese (Walker et al., 1996), Italian (Di Eugenio, 1998), Catalan and Spanish (Potau, 2008), and Romanian (Harabagiu and Maiorano, 2000).

Although there has been computational work related to Centering for generating a coherent text (Kibble and Power, 2000; Barzilay and Lee, 2004; Karamanis et al., 2009), we are not aware of any methodology or NLG system that employs ontologies to guide the generation of referential chains depending on the language considered.

3 Data collection, annotations and analysis

3.1 Material

To study the domain-specific conventions and the ways of signalling linguistic content in En-

[1] http://www.grammaticalframework.org/

glish, Swedish and Hebrew, we collected object descriptions written by native speakers of each language from digital libraries that are available through on-line museum databases. The majority of the Swedish descriptions were taken from the World Culture Museum.[2] The majority of the English descriptions were collected from the Metropolitan Museum.[3] The majority of the Hebrew descriptions were taken from Artchive.[4] Table 1 gives an overview of the three text collections. In addition, we extracted 40 parallel texts that are available under the sub-domain *Painting* from Wikipedia.

Number of	Eng.	Swe.	Heb.
Descriptions	394	386	110
Tokens	42792	27142	5690
Sentences	1877	2214	445
Tokens/sentence	24	13	13
Sentences/description	5	6	4

Table 1: Statistics of the text collections.

3.2 Syntactic annotation

All sentences in the reference material were tokenised, part-of-speech tagged, lemmatized, and parsed using open-source software. We used Hunpos, an open-source Hidden Markov Model (HMM) tagger (Halácsy et al., 2007) and Maltparser, version 1.4 (Nivre et al., 2007). The English model for tagging was downloaded from the Hunpos web page.[5] The model for Swedish was trained on the Stockholm Umeå Corpus (SUC) and is available to download from the Swedish Language Bank web page.[6] The Hebrew tagger and parsing models are described in Goldberg and Elhadad (2010).

3.3 Semantic annotation

The texts were semantically annotated by the author. The annotation schema for the semantic annotation is taken from the CIDOC Conceptual Reference Model (CRM) (Crofts et al., 2008).[7] Ten of the CIDOC-CRM concepts were employed to annotate the data semantically. These are given in table 2. Examples of semantically annotated texts are given in figure 3.[8]

Actor	Man-Made_Object
Actor Appellation	Material
Collection	Place
Dimension	Time-span
Legal Body	Title

Table 2: The semantic concepts for annotation.

3.4 Referential expressions annotation

The task of identifying referential instances of a painting entity, which is our main subject reference, requires a meaningful semantic definition of the concept *Man-Made Object*. Such a fine-grained semantic definition is available in the ontology of paintings (Dannélls, 2011),[9] which was developed in the Web Ontology Language (OWL) to allow expressing useful descriptions of paintings.[10] The ontology contains specific concepts of painting types, examples of the hierarchy of concepts that are specified in the ontology are listed below.

subClassOf(Artwork, E22_Man-Made_Object)

subClassOf(Painting, Artwork)

subClassOf(PortraitPainting, Painting and
depicts(Painting, AnimateThing))

subClassOf(OilPainting, Painting and
hasMaterial(Painting, OilPaint))

When analysing the corpus-data, we look closer at two linguistic forms of reference expressions: definite noun phrases and pronouns, focusing on three semantic relations: direct hypernym (for example *Painting* is direct hypernym of *Portrait Painting*), higher hypernym (for example, both *Artwork* and *Man-Made Object* are higher hypernyms of *Portrait Painting*) and

[2] http://collections.smvk.se/pls/vkm/rigby.welcome

[3] http://www.metmuseum.org

[4] http://www.artchive.com/

[5] http://code.google.com/p/hunpos/downloads/list

[6] http://spraakbanken.gu.se/

[7] http://cidoc.ics.forth.gr/

[8] In the Hebrew examples we use a Latin transliteration instead of the Hebrew alphabet.

[9] http://spraakdata.gu.se/svedd/painting-ontology/painting.owl

[10] http://www.w3.org/TR/owl-features/

Eng: (1) [[The Starry Night]$_{Man-Made\_Object}$]$_i$ is [[a painting]$_{Man-Made\_Object}$]$_i$ by [[Dutch Post-Impressionist artist]$_{Actor\_Appellation}$]$_j$ [[Vincent van Gogh]$_{Actor}$]$_j$. (2) Since [1941]$_{Time-Span}$ [[it]$_{Man-Made\_Object}$]$_i$ has been in the permanent collection of [the Museum of Modern Art]$_{place}$, [New York City]$_{Place}$. (3) Reproduced often, [[the painting]$_{Man-Made\_Object}$]$_i$ is widely hailed as his magnum opus.

Swe: (1) [[Stjärnenatten]$_{Man-Made\_Object}$]$_i$ är [[en målning]$_{Man-Made\_Object}$]$_i$ av [[den nederländske postimpressionistiske konstnären]$_{Actor\_Appellation}$]$_j$ [[Vincent van Gogh]$_{Actor}$]$_j$ från [1889]$_{Time-Span}$. (2) Sedan [1941]$_{Time-Span}$ har [[den]$_{Man-Made\_Object}$]$_i$ varit med i den permanenta utställningen vid [det moderna museet]$_{place}$ i [New York]$_{Place}$. (3) [[Tavlan]$_{Man-Made\_Object}$]$_i$ har allmänt hyllats som [[hans]$_{Actor}$]$_j$ magnum opus och har reproducerats många gånger och är [en av [[hans]$_{Actor}$]$_j$ mest välkända målningar]$_{Man-Made\_Object}$]$_i$.

Heb: (1) [[lila 'ohavim]$_{Man-Made\_Object}$]$_i$ hyno [[stiyor śhemen]$_{Man-Made\_Object}$]$_i$ śel [[hastayar haholandi]$_{Actor\_Appellation}$]$_j$ [[vincent van gogh]$_{Actor}$]$_j$, hametoharac lesnat [1889]$_{Time-Span}$. (2) [[hastiyor]$_{Man-Made\_Object}$]$_i$ mostag kayom [bemozehon lehomanot modernit]$_{place}$ [sebahir new york]$_{Place}$. (3) [[ho]$_{Man-Made\_Object}$]$_i$ exad hastiyorim hayedoyim beyoter sel [[van gogh]$_{Actor}$]$_j$.

Figure 3: A comprehensive semantic annotation example.

synonym, i.e. two different linguistic units of reference expressions belonging to the same concept.

3.5 Data analysis and results

The analysis consisted of two phases: (1) analyse the texts for discourse patterns, and (2) analyse the texts for patterns of coreference in the discourse.

Discourse patterns A discourse pattern (DP) is an approach to text structuring through which particular organizing principles of the texts are defined through linguistic analysis. The approach follows McKeown (1985) to formalize principles of discourse for use in a computational process. Following this approach, we have identified three discourse patterns for describing paintings that are common in the three languages. These are summarised below.

- **DP1** Man-Made_Object, Object-Type, Actor, Time-span, Place, Dimension
- **DP2** Man-Made_Object, Time-span, Object-Type, Actor, Dimension, Place
- **DP3** Man-Made_Object, Actor, Time-span, Dimension, Place

Patterns of coreference In the analysis for coreference, we only considered entities appearing in subject positions. Below follows examples of the most common types of coreference found in the corpus-data.

As seen in (1b) and in many other examples, the first reference expressions are the definite noun phrase *the painting*, i.e. coreference is build through the direct hypernym relation. The choice of the reference expression in the following sentence (1c) is the definite noun phrase *the work*, which is a higher hypernym of the main subject of reference *The Old Musician*.

(1) a. The Old Musician is an 1862 painting by French painter, Édouard Manet.
b. **The painting** shows the influence of the work of Gustave Courbet.
c. **This work** is one of Manet's largest paintings and **Ø** is now conserved at the National Gallery of Art in Washington.

Sentence (2b) shows a noun is avoided; the linguistic unit of the reference expression is a pronoun preceding a conjunction, followed by an ellipsis.

(2) a. The Birth of Venus is a painting by the French artist Alexandre Cabanel.
b. **It** was painted in 1863, and **Ø** is now in the Musée d'Orsay in Paris.

In the Swedish texts we also find occurrences of pronouns in the second sentence of the discourse, as in (3b). We learn that the most common linguistic units of the reference expressions also are definite noun phrases given by the direct hypernym relation.

(3) a. Stjärnenatten är en målning av den nederländske postimpressionistiske konstnären Vincent van Gogh från 1889.

b. Sedan 1941 har **den** varit med i den permanenta utställningen vid det moderna museet i New York.

c. **Tavlan** har allmänt hyllats som hans magnum opus och har reproducerats många gånger.

((a) The Starry Night is a painting by the dutch artist Vincent van Gogh, created in 1889. (b) Since 1941 **it** was in the permanent exhibition of the museum in New York. (c) **The picture** is widely hailed as his magnum opus and has been reproduced many times.)

Similar to English, the most common linguistic units of the reference expressions are definite noun phrases, as in (4b). However, the relation of these phrases with respect to the main subject of reference is either a direct hypernym or a synonym, such as *tavlan* in (3c) and (5b).

(4) a. Wilhelm Tells gåta är en målning av den surrealistiske konstnären Salvador Dalí.

b. **Målningen** utfördes 1933 och **Ø** finns idag på Moderna museet i Stockholm.

((a) Wilhelm Tell's Street is a painting by the artist Salvador Dali. (b) **The painting** was completed in 1933 and today it is stored in the modern museum in Stockholm.)

(5) a. Baptisterna är en målning av Gustaf Cederström från 1886, och **Ø** föreställer baptister som samlats för att förrätt dop.

b. **Tavlan** finns att beskåda i Betel folkhögskolas lokaler.

((a) The Baptists is a painting by Gustaf Cederström from 1886, and depicts baptists that have gathered for a bad. (b) **The picture** can be seen in Betel at the people's high school premises.)

The Hebrew examples also include definite noun phrases determined by the direct hypernym relation, as *hastiyor* in (6b). Pronouns only occur in a context that contains a comparison, for example (6c). In other cases, e.g. (7b), (7c), the relation selected for the reference expression is higher-hypernym.

(6) a. lila ′ohavim hyno stiyor śhemen śel hasayar haholandi vincent van gogh, hametoharac lesnat 1889.

b. **hastiyor** mosag kayom bemozehon lehomanot modernit sebahir new york.

c. **ho exad hastiyorim** hayedoyim beyoter sel van gogh.

((a) The Starry Night is an oil painting by the dutch painter Vincent van Gogh, created in 1899. (b) **The painting** is stored in the Museum of Modern Art in New York. (c) **It** is one of the most famous works of Vincent van Gogh.)

(7) a. hahalmon nehaviyon ho stiyor sel pablo picasso hametaher hames zonot.

b. **hayestira** sestzoyra ben ha sanyim 1906-1907 nehsevet lehahat min heyestirot hayedohot sel picasso vesel hahomanot hamodernit.

c. **hayestira** mosteget kayom bemostehon lehomanot modernitt sebe new york.

((a) The Young Ladies of Avignon is a painting by Pablo Picasso that portrays five prostitutes. (b) **The artwork** that was painted during 1906-1907 is one of the most known works by Picasso in the modern art. (c) **The artwork** can today be seen in the Museum of Modern Art in New York City.)

The synonym relation occurs when giving the dimensions of the painting, as in (8b).

(8) a. Soded haken (1568) ho stiyor semen al luax est meet hastayar hapalmi peter broigel haav.

b. **hatmona** hi begodel 59 al 68
centimeter, ve **Ø** motseget bemozeon
letoldot haaomanot bevina.

((a) The Nest thief (1568) is an oil painting made on wood by the painter Peter Brogel Hav. (b) **The picture** measures 59 x 68 cm, and is displayed in the art museum in Vienna.)

3.6 The results of the analysis

The above examples show a range of differences in the way chains of coreference are constructed. Table 3 summarizes the results the analysis revealed. 1st, 2nd and 3rd correspond to the first, second and third reference expression in the discourse. In summary, we found:

- Pronoun is common in Swedish and English, and rare in Hebrew
- Direct-hypernym is common in English, Swedish and Hebrew
- Higher-hypernym is rare in English and Swedish, and common in Hebrew
- Synonym is common in Swedish, less frequent in English, and rare in Hebrew

DP	English			Swedish			Hebrew		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
1	DH	P		DH	P		DH	Ø	
1	DH	HH	Ø	DH	Ø		DH		
1	P	Ø		P	Ø				
1	P	P	Ø	Ø	DH				
1				Ø	P	DH			
1,2	P	DH		P	S	Ø			
2							HH	HH	
2							HH	Ø	HH
3	P	DH		P	DH				

Table 3: Coreference strategies for a painting object realisation. Pronoun (P), Synonym (S), Direct Hypernym (DH), Higher Hypernym (HH), Ellipsis (Ø).

Although the identified strategies are constrained by a relatively simple syntax and a domain ontology, they show clear differences between the languages. As table 3 shows, consecutive pronouns occur commonly in English, while consecutive higher hypernym noun phrases are common in Hebrew.

4 Generating referential chains from Web ontology

4.1 Experimental data

We made use of the data available in the painting ontology presented in section 3.4 to generate multilingual descriptions by following the domain discourse patterns. The data consists of around 1000 ontology statements and over 250 lexicalised entities extracted from the Swedish National Museums of World Culture and the Gothenburg City Museum.

4.2 The generation grammar

The grammar was implemented in GF, a grammar formalism oriented toward multilingual grammar development and generation (Ranta, 2004). It is a logical framework based on a general treatment of syntax, rules, and proofs by means of a typed λ-calculus with dependent types (Ranta, 1994). Similar to other logical formalisms, GF separates between abstract and concrete syntaxes. The abstract syntax reflects the type theoretical part of a grammar. The concrete syntax is formulated as a set of linearization rules that can be superimposed on an abstract syntax to generate words, phrases, sentences, and texts of a desirable language. In addition, GF has an associated grammar library (Ranta, 2009); a set of parallel natural language grammars that can be used as a resource for various language processing tasks.

Our grammar consists of one abstract module that reflects the domain knowledge and is common to all languages, plus three concrete modules, one for each language, which encode the language dependent strategies. Rather than giving details of the grammatical formalism, we will show how GF captures the constraints presented in section 3.6. The examples include the following GF constructors: `mkText` (Text), `mkPhr` (Phrase), `mkS` (Sentence), `mkCl` (Clause), `mkNP` (Noun Phrase), `mkVP` (Verb Phrase), `mkAdv` (Verb Phrase modifying adverb), `passiveVP` (Passive Verb Phrase), `mkN` (Noun).

English

```
painting paintingtype painter
              year museum = let
str1 : Phr = mkPhr
(mkS (mkCl (mkNP painting) (mkVP
(mkVP  (mkNP
(mkNP a_Art paintingtype) make_V2))
(mkAdv by8agent_Prep
(mkNP (mkNP  painter)
(mkAdv in_Prep year.s)))))));
str2 : Phr = mkPhr (mkS
(mkCl (mkNP the_Art paintingtype)
(mkVP (passiveVP display_V2)
(mkAdv at_Prep  museum.s))))
in mkText str1 (mkText str2) ;
```

Swedish

```
painting paintingtype painter
              year museum = let
str1 : Phr = mkPhr
(mkS (mkCl (mkNP  painting)
(mkVP (mkVP
(mkNP a_Art paintingtype))
(mkAdv by8agent_Prep
(mkNP (mkNP painter)
(mkAdv from_Prep (mkNP year)))))));
str2 :  Phr = mkPhr
(mkS (mkCl (mkNP the_Art
(mkN "tavla" "tavla"))
(mkVP (mkVP (depV  finna_V))
(mkAdv on_Prep (mkNP museum)))) )
in mkText str1 (mkText str2) ;
```

Hebrew

```
painting paintingtype painter
              year museum = let
str1 : Str = ({s =  painting.s ++
paintingtype.s ++  "sl " ++
painter.s ++ "msnt " ++ year.s}).s;
str2 : Str = ({s = artwork_N.s ++
(displayed_V ! Fem) ++ at_Prep.s ++
museum.s}).s in
ss (str1 ++ " ." ++ str2 ++ " ." );
```

The above extracts from the concrete modules follow the observed organization principles concerning the order of semantic information in a discourse and the generation of language-dependent referential chains (presented in the right-hand column of table 4). In these extracts, variations in referential forms are captured in the noun phrase of *str2*. In the English module, the *paintingtype* that is the direct hypernym of the painting object is coded, while in the Swedish module, a synonym word of the painting concept is coded, e.g *tavla*. In the Hebrew module, a higher concept in the hierarchy of paintings, *artwork_N.s* is coded.

4.3 Experiments and results

A preliminary evaluation was conducted to test how significant is the approach of adapting language-dependent coreference strategies to produce coherent descriptions. Nine human subjects participated in the evaluation, three native speakers of each language.

The subjects were given forty object description pairs. One description containing only pronouns as the type of referring expressions and one description that was automatically generated by applying the language dependent coreference strategies. Examples of the description pairs the subjects were asked to evaluate are given in table 4. We asked the subjects to choose the description they find most coherent based on their intuitive judgements. Participant agreement was measured using the kappa statistic (Fleiss, 1971). The results of the evaluation are reported in table 5.

	Pronouns	Pronouns/NPs	$\mathcal{K}$
English	17	18	0.66
Swedish	9	29	0.78
Hebrew	6	28	0.72

Table 5: A summary of the human evaluation.

On average, the evaluators approved at least half of the automatically generated descriptions, with a considerably good agreement. A closer look at the examples where chains of pronouns were preferred revealed that these occurred in English when a description consisted of two or three sentences and the second and third sentences specified the painting dimensions or a date. In Swedish, these were preferred whenever a description consisted of two sentences. In Hebrew, the evaluators preferred a description containing a pronoun over a description containing the higher hypernym *Manmade object*, and also preferred the pronoun when a description consisted of two sentences,

English	
The Long Winter is an oil-painting by Peter Kandre from 1909. **It** is displayed in the Museum Of World Culture.	The Long Winter is an oil-painting by Peter Kandre from 1909. **The painting** is displayed in the Museum Of World Culture.
The Little White Girl is a painting by James Abbott McNeill Whistler. **It** is held in the Gotheburg Art Museum.	The Little White Girl is a painting by James Abbott McNeill Whistler. **The painting** is held in the Gotheburg Art Museum.
The Long Winter is a painting by Peter Kandre from 1909. **It** measures 102 by 43 cm. **It** is displayed in the Museum Of World Culture.	The Long Winter is a painting by Peter Kandre from 1909. **It** measures 102 by 43 cm. **The painting** is displayed in the Museum Of World Culture.
Swedish	
Den långa vintern är en oljemålning av Peter Kandre från 1909. **Den** återfinns på Världskulturmuseet.	Den långa vintern är en oljemålning av Peter Kandre från 1909. **Tavlan** återfinns på Världskulturmuseet.
Den lilla vita flickan är en målning av James Abbott McNeill Whistler. **Den** återfinns på Göteborgs Konstmuseum.	Den lilla vita flickan är en målning av James Abbott McNeill Whistler. **Målningen** återfinns på Göteborgs Konstmuseum.
Den långa vintern målades av Peter Kandre 1909. **Den** är 102 cm lång och 43 cm bred. **Den** återfinns på Världskulturmuseet.	Den långa vintern målades av Peter Kandre 1909. **Målningen** är 102 cm lång och 43 cm bred. **Tavlan** återfinns på Världskulturmuseet.
Hebrew	
hHwrP hArwK hnw Zywr smN sl pyTr qndrh msnt 1909. **hyA** mwZg bmwzAwN sl OlM htrbwt.	hHwrP hArwK hnw Zywr smN sl pyTr qndrh msnt 1909. **hZywr** mwZg bmwzAwN sl OlM htrbwt.
hyaldh hktnh alevmh hi tmona sl abut mcnil wistl. **hyA** mwZgt bmwzAwN homanot sl gwTnbwrg.	hyaldh hktnh alevmh hi tmona sl abut mcnil wistl. **hyZyrh** mwZgt bmwzAwN homanot sl gwTnbwrg.
HwrP ArwK tzoyar el–yedy pyTr qndrh b–1909. **hyA** bgwdl 102 Ol 43 Sg2m. **hyA** mwZgt bmwzAwN sl OlM htrbwt.	HwrP ArwK tzoyar el–yedy pyTr qndrh b–1909. **hyZyrh** bgwdl 102 Ol 43 Sg2m. **hyZyrh** mwZgt bmwzAwN sl OlM htrbwt.

Table 4: Examples of object description pairs that were used in the evaluation.

the second of which concerned the painting dimensions.

5 Conclusions and future work

This paper has presented a cross-linguistic study and demonstrated some differences in how coreference is expressed in English, Swedish and Hebrew. As a result of the investigation, a set of language-specific coreference strategies were identified and implemented in GF. This multilingual grammar was used to generate object descriptions which were then evaluated by native speakers of each language. The evaluation results, although performed with a small number of descriptions and human evaluators, indicate that language-dependent coreference strategies lead to better output. Although the data used to compare the co-referential chains was restricted in size, it was sufficient to determine several differences between the languages for the given domain.

Future work aims to extend the grammar to cover more ontology statements and discourse patterns. We will consider conjunctions and ellipsis in these patterns. We intend to formalize and generalize the strategies presented in this paper and test whether there exist universal coreferential chains, which might result in coherent descriptions in more than three languages.

Acknowledgments

The research presented in this paper was supported in part by MOLTO European Union Seventh Framework Programme (FP7/2007-

2013) under grant agreement FP7-ICT-247914.[11] I would like to thank the Centre for Language Technology (CLT) in Gothenburg and the anonymous INLG reviewers.[12]

References

Mira Ariel. 1990. *Accessing Noun Phrase Antecedents*. Routlege, London.

Regina Barzilay and Lillia Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proc. of HLT-NAACL*, pages 113–120.

Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff, 2008. *Definition of the CIDOC Conceptual Reference Model*.

Dana Dannélls. 2011. An ontology model of paintings. *Journal of Applied Ontologies*. Submitted.

B. Di Eugenio, 1998. *Centering in Italian*, pages 115–137. Oxford: Clarendon Press.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. 2008. Attempto Controlled English for Knowledge Representation. In *Reasoning Web, Fourth International Summer School*. Springer.

T. Givón, editor. 1983. *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam and Philadelphia: John Benjamins.

Yoav Goldberg and Michael Elhadad. 2010. An efficient algorithm for easy-first non-directional dependency parsing. In *Proc. of NAACL 2010*.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2).

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos: an open source trigram tagger. In *Proc. of ACL on Interactive Poster and Demonstration Sessions*, pages 209–212, Morristown, NJ, USA.

Michael A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman Pub Group.

S. Harabagiu and S. Maiorano. 2000. Multilingual coreference resolution. In *Proc. of ANLP*.

Anna Sågvall Hein. 1989. Definite NPs and background knowledge in medical text. *Computer and Artificial Intelligence*, 8(6):547–563.

Nikiforos Karamanis, Massimo Poesio, Chris Mellish, and Jon Oberlander. 2009. Evaluating Centering for Information Ordering using Corpora. *Computational Linguistics*, 35(1).

Rodger Kibble and Richard Power. 2000. Optimizing Referential Coherence in Text Generation. *Computational Linguistics*, 30(4).

J. McCrae, G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gomez-Perez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*.

Kathleen R. McKeown. 1985. *Text generation : using discourse strategies and focus constraints to generate natural language text*. Cambridge University Press.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Marta Recasens Potau. 2008. *Towards Coreference Resolution for Catalan and Spanish*. Ph.D. thesis, University of Barcelona.

Rashmi Prasad. 2003. *Constraints on the generation of referring expressions, with special reference to hindi.* Ph.D. thesis, University of Pennsylvania.

Ellen F. Prince. 1992. The ZPG letter: Subjects, definiteness, and information-status. In *Discourse description. diverse linguistic analyses of a fund-raising text*, volume 10, pages 159–173.

Aarne Ranta. 1994. *Type-theoretical grammar: A Type-theoretical Grammar Formalism*. Oxford University Press, Oxford, UK.

Aarne Ranta. 2004. Grammatical Framework, a type-theoretical grammar formalism. *Journal of Functional Programming*, 14(2):145–189.

Aarne Ranta. 2009. The GF resource grammar library. *The on-line journal Linguistics in Language Technology (LiLT)*, 2(2).

R. Schwitter and M. Tilbrook. 2004. Controlled Natural Language meets the Semantic Web. In *Proceedings of the Australasian Language Technology Workshop*, pages 55–62, Macquarie University.

Enric Vallduví and Elisabet Engdahl. 1996. The linguistic realization of information packaging. *Linguistics*, (34):459–519.

M. A. Walker, M. Iida, and S. Cote. 1996. Centering in Japanese Discourse. *Computational Linguistics*.

Sandra Williams, Allan Third, and Richard Power. 2011. Levels of organisation in ontology verbalisation. In *Proc. of ENLG*, pages 158–163.

[11] `http://www.molto-project.eu/`

[12] `http://www.clt.gu.se/`