ISSN 1744-1986

Technical Report N° 2013/ 02

An Analysis of POS Tag Patterns in Ontology Identifiers and Labels

Sandra Williams

, 2013

Department of Computing
Faculty of Mathematics, Computing and Technology
The Open University

Walton Hall, Milton Keynes, MK7 6AA
United Kingdom

http://computing.open.ac.uk

An Analysis of POS Tag Patterns in Ontology Identifiers and Labels

Sandra Williams
The Open University, Milton Keynes, MK7 6AA, U.K.

Abstract

I describe an analysis of the syntax of identifier names found in a corpus of over 500 ontologies.[1] The analysis was performed in five steps: (i) extraction of identifier names from the corpus; (ii) construction of dummy sentences containing the identifiers; (iii) part-of-speech (POS) tagging; (iv) extraction of POS tag strings; (v) POS string frequency analysis; and (vi) general syntactic pattern analysis. The findings of the analysis were that identifier names follow simple syntactic patterns; each type of identifier can be expressed through relatively few patterns; and the syntax of identifiers differs from natural English in consistent ways.

1 Introduction

Ontologies are a means of representing human knowledge. They are employed extensively in the semantic web where they are expressed in ontology languages such as Web Ontology Language (OWL) or RDF triples. Underlying OWL and RDF are simple, but powerful, description logics, which are expressed in statements (axioms) such as: $A \sqsubseteq B$ (A is a subclass of B, e.g., every man is a human), $A \sqsubseteq \exists P.B$ (A is a subclass of things that have property P (relationship) with B, e.g., every pizza is covered in cheese), $[a, b] \in P$ ('a' and 'b' are individuals related by the property P, e.g., John loves Mary), and $a \in A$ ('a' is a member of class A, e.g., John is a man).

The analysis described in this paper was carried out to determine the syntactic structures of ontology identifier names, free text labels created by human authors when they construct an ontology. The aim of this study was to find out whether identifier names follow common syntactic patterns; and whether the patterns differ from those present in natural language (English). Part-of-speech (POS) tag strings were computed from the identifier names present in a large corpus of ontologies, for example, NN IN JJ NN from

[1]This research was carried out under the Semantic Web Authoring Tool (SWAT) project funded by the U.K. Engineering and Physical Science Research Council (EPSRC grant no. G033579/1).

Table 1: Penn part-of-speech tagset

Tag	Description	Tag	Description
CC	conjunction, coordinating	PRP$	pronoun, possessive
CD	numeral, cardinal	RB	adverb
DT	determiner	RBR	adverb, comparative
EX	existential there	RBS	adverb, superlative
FW	foreign word	RP	particle
IN	preposition, conjunction	SYM	symbol
JJ	adjective or numeral, ordinal	TO	'to' prep. or infin. marker
JJR	adjective, comparative	UH	interjection
JJS	adjective, superlative	VB	verb, base form
LS	List item marker	VBD	verb, past tense
MD	modal auxiliary	VBG	verb, pres. participle, gerund
NN	noun, common, sing. or mass	VBN	verb, past participle
NNP	noun, proper, sing.	VBP	verb, pres. tense, not 3rd sing.
NNPS	noun, proper, plural	VBZ	verb, pres. tense, 3rd sing.
NNS	noun, common, plural	WDT	WH-determiner
PDT	pre-determiner	WP	WH-pronoun
POS	genitive marker	WP$	WH-pronoun, possessive
PRP	pronoun, personal	WRB	Wh-adverb

'ZoneOfLacrimalBone', where NN is the tag for singular common nouns 'zone' and 'bone', IN is for prepositions such as 'of', and JJ is for adjectives such as 'lacrimal'. A further goal was to to develop syntactic patterns from common POS-tag strings to be used for automatically extracting candidate identifier names from free text as an aid to ontology developers.

2 Part-of-Speech Tagging

Part-of-speech tagging is the task of annotating a text with parts of speech. This study used the Stanford POS tagger[2] (Toutanova et al., 2003) which is ready-trained on large corpora and uses the Penn tagset (Santorini, 1990) reproduced in table 1. Developers of the Stanford POS tagger claim a token accuracy of over 97% (i.e., two or three mis-tagged words per 100) for test texts from the same corpus as the texts on which it was trained. Accuracy for unknown words is claimed at over 80%. It is easily deployed within a Java program by implementing an external system call (in effect, the same as calling it from the command line in a Linux shell).

Apart from mis-tags, other disadvantages of the tagger are that it is slow and it requires a lot of memory, especially for texts of 1MB or more (I allocated 2GB of RAM in the system call).

[2] I used the December 2011 version of the Stanford POS tagger, downloaded from http://nlp.stanford.edu/software/tagger.shtml in May 2012

3 Method

Identifier names were extracted as phrases from ontology files in a corpus of 521 ontologies using one of the SWAT Natural Language tools[3] which strips off the namespace, replaces aliases with the definition of the alias (thus eliminating abbreviated IRIs), uses labels where atomic terms are annotated with them, and splits the remaining part of the identifier or label into words using underline characters or camel-case as delimiters. Four large files were extracted, one for each of the following types of identifier: classes, named individuals, object properties and data properties. All except named individuals were converted to lower case. Each type of identifier was then inserted into a dummy sentence before POS tagging because the POS tagger is trained on complete sentences and therefore would not work well on phrase fragments. A previous, smaller study had confirmed our assumptions that class names and named individuals tend to be nouns and proper nouns, respectively, and also that property names tend to be verbs (Power, 2010). I therefore positioned the identifier names in dummy sentences as noun, proper noun and verbs, as follows:

- Every <class_name> is a thing.
- <Individual_name> is a thing.
- Everything <object_property_name> a thing.
- Everything <data_property_name> a thing.

POS tagging with the Stanford POS tagger was then performed. Tagged identifiers were extracted from the tagged dummy sentences, and printed one per line in a file (according to type) with identifier string first, followed by a field delimiter (&), and finally the POS tag string, e.g., 'water table & NN NN'. This file could then be loaded into a spreadsheet application[4] with identifier strings in the first column and POS tag strings in the second. The spreadsheet was sorted by POS tag string, and subtotals for each unique POS tag string calculated. Results for each identifier type are presented in the following section.

4 Results

4.1 Results of the Class Identifier Analysis

From the corpus of 521 ontologies, 165,930 class identifiers were extracted. After removal of duplicate identifiers and obvious numerical (non-verbal) identifiers such as 'gene23456', 62788 human-authored identifiers remained

[3] http://swat.open.ac.uk/tools/

[4] I used Microsoft Excel.

Table 2: TOP 25 CLASS IDENTIFIER POS TAG PATTERNS Identifier names are converted to lower case and obvious numerical labels such as ‘emap 377’ eliminated. Percentages are rounded up or down to the nearest whole number.

	Tag String	**Example**	**Freq**	**%**
1	NN	calcium	14718	23%
2	NN NN	water table	13384	21%
3	JJ NN	electronic noise	8631	14%
4	JJ NN NN	periodic plane tesselation	1640	3%
5	NNS	neurons	1458	2%
6	NN NNS	life jackets	1372	2%
7	NN NN NN	organ rejection process	1338	2%
8	JJ JJ NN	acute gastric ulcer	1007	2%
9	JJ NNS	fluid dynamics	962	2%
10	NN IN NN	calix of kidney	772	1%
11	FW FW	exerpes asper	524	1%
12	NN IN JJ NN	zone of lacrimal bone	519	1%
13	JJ NN IN JJ NN	bony part of right fibula	490	1%
16	JJ	antimicrobial	464	1%
17	VBP	fluctuate	425	1%
18	JJ NN IN NN	upper lobe of lung	386	1%
19	JJ NN NN NN	distal tip cell migration	386	1%
20	NNS NN	species diversity	375	1%
21	JJ NN IN JJ NN IN JJ JJ NN	distal epiphysis of distal phalanx of left big toe	287	0%
22	NN IN NN NN	moment of force unit	278	0%
23	NN IN NN CD NN	mole of radium 2 ion	270	0%
24	NN IN JJ JJ NN	tubercle of right third rib	244	0%
25	VBG	preventing	233	0%
		Total (above patterns)	**50163**	80%
		Grand Total (class identifiers)	**62788**	100%

for analysis. Frequency counts for the top 25 class identifiers are shown in table 2.

Table 2 demonstrates that surprisingly few POS tag strings account for the majority of class identifiers; indeed, those shown in the table account for 80%. Further analysis of the strings revealed that *only one* syntactic pattern covers 77% of all identifier names for classes in the corpus (including POS tag strings in the remaining 20% of the corpus not shown in table 2):

- **Pattern 1:** $<\text{JJ}>^{*}\ <\text{NNS?}>^{+}\ <<\text{IN}>^{+}\ <\text{JJ}>^{*}\ <\text{NNS?}>^{+}>^{*}$

where the symbols used are as shown in table 3. Adding a further two patterns brings the coverage up to 91%:

- **Pattern 2:** $<\text{FW}>^{+}$
- **Pattern 3:** $<\text{JJ}>^{*}\ <\text{NNS?}>^{+}\ <\text{CD}>^{+}$

Table 3: Symbols

$<>$	delimit an element
$*$	zero or more occurrences, e.g., $< JJ > *$
$+$	one or more occurrences, e.g., $< NN > +$
$\mid$	alternatives, e.g., IN$\mid$TO
?	the previous character is optional, e.g., NNS?

However, pattern 1 is by far the most important. Patterns 2 and 3 are less important; pattern 3 does not even occur in the table, although it is fairly common amongst the POS tag strings not shown in the table.

As noted by Power (2010), there are some essential similarities and differences between class identifiers and natural English noun phrases. Class identifiers can be characterised by the following criteria:

- Definite and indefinite articles (DT) are not included.
- Overwhelmingly, the commonest patterns consist of common nouns, NN and NNS, adjectives, JJ, and prepositions, IN. Similar to patterns for noun phrases and prepositional phrases in English, except that these patterns omit articles (DT).
- Many patterns (but not the commonest) contain foreign words, FW, or cardinal numbers, CD. These are often technical terms such as the Latin 'exerpes asper' and the NN CD pattern 'radium 2'.
- Verbs are rare (see VBP and VBG in table 2).
- Proper nouns, NNP and NNPS, are not included.
- If there is no preposition, the final common noun is usually the head which should be modified to form the plural, and might indicate a superclass, e.g., 'upper lobe' is a 'lobe'
- If prepositions are present, the common noun before the final preposition is often the head noun which should be modified to form the plural and might also indicate a superclass, e.g., 'upper lobe of lung' is a 'lobe' .

4.2 Results of the Named Individual Identifier Analysis

POS tag pattern analysis of named individual identifiers was carried out in a similar way to that of class identifiers except that the identifiers were *not* converted to lower case. I removed around 10,000 obvious numerical identifiers analysing only the remaining 13,598 identifiers. The results, shown in table 4, are more varied than for class identifiers.

Table 4: TOP 25 NAMED INDIVIDUAL POS TAG PATTERNS Obvious numerical labels such as 'emap 377' were eliminated. Percentages are rounded up or down to the nearest whole number.

	Tag String	**Example**	**Freq**	**%**
1	NNP	Abruzzo	2585	19%
2	NN	soil	1828	13%
3	NNP NNP	Ynys Moelfre	1426	10%
4	NN NN	Nucleus Type	713	5%
5	NNP NNP NNP	Pen Y Ghent	431	3%
6	JJ NN	Primary Culture	418	3%
7	NNS	Paints	315	2%
8	JJ NN NN	Axonal Growth Cone	223	2%
9	NNP NNP CD	Sarah Rever 1850	206	2%
10	NNP NN	Whitehall Lane	170	1%
11	NNP NNP NNP CD	Mary Ann Green 1846	144	1%
12	FW FW FW	Realschule Bayern 10te	134	1%
13	JJ NNS	Olfactory organs	121	1%
14	NNP NNP NNP NNP	Geo Protocol Google Maps	113	1%
15	NN NNS	Use tools	110	1%
16	NN NN NN	Page Mill Winery	107	1%
17	VB DT CD	maskge a 3	105	1%
18	JJ	Spiny	94	1%
19	NNP NNS	Small Isles	89	1%
20	NNP NNP NN	Albania Greece border	74	1%
21	NN NN CD	john cotton 1778	72	1%
22	NN DT CD	shiftdata a 4	66	0%
23	NNP CD	Fouriertransform 1	59	0%
24	VBG	Melting	57	0%
25	NNP NNS	Clay minerals	54	0%
		Total (above patterns)	**9714**	71%
		Grand Total (named individual identifiers)	**13598**	100%

Further analysis revealed that the majority of these identifiers, not surprisingly, contain one or more singular or plural proper nouns (5351 of the 9714 in the table, or 55%). Further analysis of the strings revealed that *three* syntactic patterns cover 43% of all POS tag strings derived from named individual identifier names; these are:

- **Pattern 4:** <JJ>* <NNS?>* <NNPS?>$^+$ <NNS?>*
- **Pattern 5:** <JJ>* <NNS?>* <NNPS?>$^+$ <NNS?>* <CD>$^+$
- **Pattern 6:** <NNS?>* <NNPS?>$^+$ <IN>$^+$ <NNS?>$^+$

Note that pattern 1 for class identifiers would account for many POS tag strings derived from named individual identifiers (e.g., those in rows 2, 4, 6, 7, 8, 13, 15 and 16 in table 4). I also observed the following characteristics some of which demonstrate similarities and differences between syntactic patterns for named individuals and those of English proper names:

- As with class names, definite and indefinite articles, DT, are not included apart from mis-tags such as ‘shiftdata a 4’ in row 22 in the table.
- Prepositions, IN, are not included.
- Verbs are rarely included (apart from VBG in row 24).
- The majority consist of one or more proper nouns, NNP and NNPS (5351 of 9714 in the table, or 55%).
- Often, numbers are used to distinguish non-unique names or some property (e.g., birth date as in ‘Mary Ann Green 1846’ in row 11).
- Common nouns, NN and NNS are often mis-tagged as proper nouns, NNP and NNPS (e.g., ‘Maps’ in row 14), and vice versa (e.g., ‘john cotton’ in row 21); capitalisation seems to influence tagging.
- Common patterns are four words or fewer in length.

These findings are similar to Power’s smaller analysis (Power, 2010), except that he found prepositions occurred in 1% of the patterns he analysed in a smaller corpus, e.g., ‘Bay of Biscay’ with a frequency of 33, however, the pattern was much less frequent in this larger corpus analysis (less than 0.4%).

4.3 Results of the Object Property Identifier Analysis

POS tag pattern analysis was carried out as before. The results in table 5 show fewer identifiers of this type in the corpus and greater diversity than in class and named individual identifiers. They also demonstrate that human authors frequently construct property names from relational combinations of noun + preposition (e.g., ‘locus of’), even though verbs would be the more obvious choice to express the role of object properties in defining relationships between named individuals in ontologies.

Six syntactic patterns cover 59% of POS tag strings derived from object property identifier names; these are:

- **Pattern 7:** <VBZ>$^{+}$ <JJS?>* <NNS?>$^{+}$ <IN>*
- **Pattern 8:** <VBZ>* <VBN>$^{+}$ <NNS?>* <IN|TO>*
- **Pattern 9:** <JJ>* <NNS?>$^{+}$ <IN>$^{+}$
- **Pattern 10:** <NNS?>* <VBD>$^{+}$ <NNS?>* <IN|TO>*

Table 5: TOP 25 OBJECT PROPERTY POS TAG PATTERNS Percentages are rounded up or down to the nearest whole number. *Mis-tagged.

	Tag String	**Example**	**Freq**	**%**
1	VBZ NN	has storey	768	9%
2	NN	height	601	7%
3	VBZ NN IN	is provider of	424	5%
4	VBZ NN NN	has risk type	400	4%
5	NN NN	year value	390	4%
6	VBZ JJ NN	has great uncle	333	4%
7	VBZ	eats	330	4%
8	NN IN	locus of	306	3%
9	VBN IN	offered at	233	3%
10	VBZ VBN IN	is imported by	223	3%
11	JJ NN	standard error	216	2%
12	VBZ JJ NN IN	is physical state of	181	2%
13	VBD	surname*	162	2%
14	VBZ VBN	has mimetype*	156	2%
15	NN NN IN	body attitude of	127	1%
16	VBZ NN NN IN	is review history of	124	1%
17	NN VBP	domain attribute*	103	1%
18	NN VBD	location id*	100	1%
19	NN NN VBD	rel point ref*	94	1%
20	JJ NN IN	distinct part of	93	1%
21	NN NN NN	cargo type code	88	1%
22	NN IN NN	day of year	74	1%
23	VBZ JJ JJ NN IN	is medial lateral selector of	72	1%
24	VBZ IN	starts during	67	1%
25	VBZ VB	has gender*	67	1%
		Total (above patterns)	**5732**	64%
		Grand Total (object property identifiers)	**8918**	100%

- **Pattern 11:** <JJ>* <NNS?>* <VBP>$^+$
- **Pattern 12:** <VB>$^+$

Note that, once again, pattern 1 for class identifiers would account for many object property identifiers too (rows 2, 5, 11, 21 and 22 in table 5). Clearly ontology authors often omit verbs such as 'is' or 'has' and prepositions from object property names. However, this creates a problem of readability since someone reading the ontology cannot tell whether a property such as the bare noun 'value' should mean 'has a value of' or 'is the value of'.

I observed the following characteristics that demonstrate similarities and differences from natural English:

- The majority are phrases composed from a noun, often followed by a preposition, optionally with 'is' or 'has' at the beginning; e.g., see 'is imported by' in row 10. Indeed, where the tag VBZ occurs, it usually tags 'has' or 'is'.
- Many are composed entirely of common nouns (see rows 1, 3, 5, 6, 8, and 17 in the table). Some nouns have been mis-tagged as verbs (see rows 13, 14, 17, 18, 19, and 25 in the table); this may be a side-effect of the placement of object property identifiers in dummy sentences.
- There are fewer lone transitive verbs than expected, e.g., row 7, and many of these are mis-tagged nouns.
- As before, there are no definite or indefinite articles, DT.

4.4 Results of the Data Property Identifier Analysis

The results in table 6 show that there are few of this identifier type in the corpus (only 3501 in total). *Three* syntactic patterns cover 60% of POS tag strings derived from data property identifier names; however they are identical to pattern 1 for classes and patterns 7 and 10 for object properties.

Observations about their characteristic similarities and difference from natural English are similar:

- As with the other types, no determiners are present.
- As with object property identifiers, greater proportions are based on nouns than verbs (fewer than a third contain verbs).
- Unlike object properties, prepositions are rare (see 'date of birth' in row 19).
- Many (around 6%) are longer than other identifier types, consisting of five words or more (see rows 8, 17, 22, and 24 in the table).
- Many are mis-tagged (see asterisked entries in the table).

Table 6: TOP 25 DATA PROPERTY POS TAG PATTERNS Percentages are rounded up or down to the nearest whole number. *Mis-tagged.

	Tag String	Example	Freq	
1	NN NN	axiom content	368	11%
2	NN	value	294	8%
3	NN NN VBD	bridge type id*	202	6%
4	NN VBD	arg desc*	166	5%
5	NN NN NN	time span end	142	4%
6	NN NN NN NN	convoy day speed rate	133	4%
7	VBD	password*	133	4%
8	NN NN NN NN NN	fan area sector size angle	109	3%
9	VBZ NN	has currency	102	3%
10	VBZ NN NN	has cost amount	87	2%
11	JJ NN	first name	79	2%
12	VBZ JJ NN	has raw string	73	2%
13	NN NN NN VBD	feat lock stat id*	70	2%
14	NNS VBP	pref label*	56	2%
15	VBZ VBN	is ordered	52	1%
16	JJ NN NN	postal code txt	35	1%
17	NN NN NN NN NN NN	network service access call sign text	33	1%
18	VBZ NN NN NN	has data service type	30	1%
19	NN IN NN	date of birth	26	1%
20	NN VBP	medium note*	26	1%
21	NN NN VBP	safety stat code*	23	1%
22	NN NN NN NN VBD	point ref orgn point id*	21	1%
23	VBZ JJ	is infinite	21	1%
24	NN NN JJ NN NN	crypto plan short title txt	19	1%
25	VBD NN	translated title	19	1%
		Total (above patterns)	**2319**	67%
		Grand Total (data property identifiers)	**3501**	100%

5 Related work

Apart from the smaller analysis of Power (2010), two other studies exist, Sun and Mellish's (2007) and Marco Trevisan's (2010). Directly comparing proportions of *class* POS tag strings with Sun and Mellish and Power[5] is possible (but not Trevisan because he did not analyse class identifiers); the comparison of class identifier POS tag strings is shown in table 7 where only high frequency POS tag strings that occur in more than one corpus are shown. Agreement over the top POS tag strings is fairly good (rows 1–4 in the table) although Mellish and Sun's most frequent patterns were proportionally smaller (their top four patterns consisted of only 35% of class names in their corpus). Sun and Mellish mention that 72% of POS tag

[5]Power's analysis was conducted on a subset of our corpus

Table 7: Comparson of frequencies of class identifier POS tag strings

POS Tag String	Mellish and Sun's	Power's Corpus	Our Corpus
NN	14%	16%	23%
NN NN	11%	17%	21%
JJ NN	4%	22%	14%
JJ NN NN	2%	7%	3%
NN NN NN	4%	5%	2%
JJ JJ NN	-	4%	2%
Total	**37260**	**12437**	**62788**

strings for class names from their corpus end with nouns; this agrees well with Power (around 80%) and our analysis (around 80%). Sun and Mellish also observed that 30% of POS tags strings for class names in their corpus consist entirely of nouns; Power's corpus had a greater proportion (around 40%), but our corpus contained an even higher proportion of noun-only strings (between 50% and 60%).

Table 8: Comparson of frequencies of property identifier POS tag strings.

POS Tag String	Mellish and Sun's Corpus	Trevisan's Corpus	Power's Corpus	Our Corpus
VBZ NN	5%	2%	26%	9%
NN	6%	4%	3%	7%
VBZ NN IN	1%			5%
NN NN		10%	3%	4%
VBZ JJ NN			10%	4%
VBZ			4%	4%
VBD	10%	1%		2%
NN IN	3%		19%	3%
VBN IN	5%	1%	3%	3%
VBZ VBN IN	3%			3%
VBZ VBN	10%			2%
NN NN IN			6%	1%
Total	**1354**	**???**	**413**	**5732**

In table 8 we compare relative frequencies of POS tag strings for *object property identifiers* from our corpus with *object and data property identifiers* from Sun and Mellish's and Power's corpus. Some results from Trevisan's corpus are also included in the table; it is not clear whether his results include both object and data property identifiers nor is it clear how many of these were present in his corpus of 200 ontologies.[6]

[6]Treviso mentions that he analysed over 12000 identifiers but did not give figures by type.

Overall, there is little agreement between the analyses. Trevisan observed that there was little overlap between POS tag strings in his corpus and those in Sun and Mellish's corpus, in particular, Sun and Mellish's patterns contained more verbs than Trevisan's. Our results differ from the others possibly because in the table we only include object property proportions. Property identifiers are highly variable, however, our more-detailed analysis of the differences between object and data property names in sections 4.3 and 4.4 demonstrated that object properties are expressed through a greater number of semantic patterns than data properties; perhaps results were compared for each of these types across all the corpora, they might exhibit greater agreement.

6 Conclusions

The aim of this study was to find out whether identifier names follow common syntactic patterns; and whether the patterns differ from those present in natural language (English). My analyses of the four types has shown that they do indeed follow common syntactic patterns; in each case, a very small set of three to six patterns accounts for a large proportion of POS tag strings. They demonstrate clear similarities and differences from natural English and can be best described as a shorthand, or telegraphic, form of English.

Class names are similar to short English noun phrases that may, in addition, have one or two postmodifiers similar to English prepositional phrases. The major difference is that the phrases do not contain determiners; where, in English, we would write 'the upper lobe of the lung', the class identifier is 'upper lobe of lung'. Another difference is that whereas in English we might expect a noun phrase describing a class of things to be plural, e.g., 'life jackets', the majority of ontology authors use singular class names.

Named individual identifiers are similar to English noun phrases and English proper names. Like class names, the major difference is the lack of determiners. An additional difference is a lack of prepositions; in English, we would expect names such as 'the leader of the opposition' but these are rare amongst named individual identifier names.

Object and data property identifier names are not often expressed through English verbs, as we would have expected from their role in ontologies. Instead, their form is/has + noun + preposition, e.g., 'is provider of'. The major differences between the two types are that prepositions are more common in object property names and there is greater variation in their patterns; data property name patterns overlap substantially with the major first pattern of class names. In both types, whilst a noun is nearly always present, 'is' or 'has' or a preposition are frequently omitted ('has amount', 'provider of' or 'provider'); a bare noun such as 'provider' makes the meaning ambiguous ('is provider' or 'has provider'?).

References

Power, Richard. 2010. Analysis of identifiers and labels in ontologies. Technical Report Unpublished.

Santorini, Beatrice. 1990. Part-of-speech tagging guidelines for the penn treebank project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Sun, Xiantang and Chris Mellish. 2007. Domain independent sentence generation from rdf represent ations for the semantic web. In *ECAI'06 the Eleventh European Workshop on Natural Language Generation, Association for Computational Linguistics*, pages 105–108, Stroudsburg, PA, USA.

Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pages 252–259.

Trevisan, Marco. 2010. A Portable Menuguided Natural Language Interface to Knowledge Bases for Querytool. Technical Report KRDB10-01, KRDB Research Centre for Knowledge and Data, Faculty of Computer Science, Free University of Bozen-Bolzano.