Linguistic Information and Standards in uComp Resources

Wim Peters
Department of Computer Science
University of Sheffield

1. Introduction

Individual linguistic and terminological resources, of the kind enabled and produced by means of uComp human computation technology, greatly differ in the explicit linguistic information they capture, which may vary in format, content granularity and the motivation for their creation, such as the immediate needs of the intended user. For instance, if we focus on the main area of linguistic description covered by this deliverable, part of speech information for English output, for one user it could be sufficient to use part of speech tags at a general level (e.g. “noun”), while for another more specific information is necessary (e.g. singular noun).

In order to accommodate these factors we need interoperability between linguistic information from heterogeneous sources formats and levels of description. This requires a networking of tag sets and vocabularies, and adherence to standards for linguistic description.

2. Standards for the representation of linguistic information

There are a number of initiatives to make conceptual and linguistic classifications interoperable and exploitable in a uniform fashion. This has resulted in various (established/proposed/de facto) standards and best practices for encoding linguistic and terminological knowledge, both from the (computational) linguistic and the semantic web side. These differ in representation format and level of formalization. For instance, many linguistic resources such as text corpora, thesauri and dictionaries are encoded in XML, but an increasing number of linguistic resources are represented as populated RDF or Owl models in order to be exploitable in semantic web applications.

Standards and best practises include the Lexical Markup Framework (LMF) [7], which presents a linguistic description of lexical knowledge, whereas Lemon [8] is also a model for sharing lexical information on the semantic web.

GOLD[1] [9] is a richly axiomatized ontology for descriptive linguistics. It is intended to capture the knowledge of a well-trained linguist, and can thus be viewed as an attempt to codify the general knowledge of the field.

With respect to interoperability, The NLP Interchange Format[2] (NIF) is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations.

OLiA[3] represents a repository of reference categories for morphosyntax, syntax and is informally interlinked with ISOCAT and GOLD. In the translation memory area, standards

[1] http://linguistics-ontology.org/
[2] http://nlp2rdf.org/nif-1-0
[3] http://nachhalt.sfb632.uni-potsdam.de/owl/

such as Translation Memory eXchange[4] and the XML Localization Interchange File Format (XLIFF)[5] are widely used.

CLARIN[6] is committed to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at offering a stable, persistent, accessible and extendable eHumanities infrastructure. META-NET[7] is a Network of Excellence serving the multilingual European information society by establishing interoperability between language technology and resources. A recently established W3C Ontolex interest group[8] is developing a model for lexicons and the relation of lexical meaning with ontologies, and investigates the added value of using such a model in semantic web NLP applications. The Open Linguistics Working Group of the Open Knowledge Foundation[9] works towards a linked open data cloud of linguistic resources, which applies the linked data paradigm to linguistic knowledge.

3. Interoperability

Although at present there are many converging developments, the picture is still diverse, and the user must choose between standards, which complement each other or overlap to a certain extent. Moreover, there are many non-standard models with deviating terminology and coverage compound the linguistic confusion.
Therefore, given the existence of this variety of (standard) linguistic models, it is necessary to establish interoperability between their vocabularies in a principled way in order to enable interdisciplinary re-use and comparison.
One initiative in this direction is the ISOCAT data category registry[10] [4], which enables the linking of elements from different linguistic data category sets. The linking facility RELCAT [3] defines a number of relations in order to accommodate the linking of local/personal linguistic data categories to elements from the ISOCAT registries, and bears resemblance to SKOS[11]. Furthermore, [10] describes an Owl-based mapping model for establishing links between descriptive vocabulary elements.

Once complete interoperability will have been established between linguistic representation formats and their content, the full range of linguistic description will be available for exploitation within a text mining infrastructure.

4. Standardization of uComp part of speech information

In order to maximize the interoperability of the part of speech annotations provided by UComp, a GATE pipeline has been created which adds various notations derived from

---

[4]
[5] http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.pdf
[6] http://www.clarin.eu
[7] http://www.meta-net.eu/
[8] http://www.w3.org/community/ontolex/
[9] http://okfn.org/
[10] http://www.isocat.org/
[11] http://www.w3.org/TR/2009/REC-skos-reference-20090818/

widely used de facto standards/best practice descriptors. They all pertain to the English language.

The following tag sets are included in the output of the GATE English part of speech annotation pipeline.

1) Penn Treebank[12]
This is one of the most widely used sets and consists of 36 tags[13] at a medium level of granularity [7].

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle

2) ISOCAT
The ISOCAT data category registry[14] [8] covers a ISO standardized set of canonical linguistic data types. It enables the linking of elements from different linguistic data category sets by means of the linking facility RELCAT [3], which defines a number of relations in order to accommodate the linking of local/personal linguistic data categories to elements from the ISOCAT registries. Its linking relations bear resemblance to SKOS[15].

3) Universal POS tags[16]

[12] http://www.cis.upenn.edu/~treebank/
[13] https://mlnl.net/jg/software/pac/ptb_pos.html
[14] http://www.isocat.org/
[15] http://www.w3.org/TR/2009/REC-skos-reference-20090818/
[16] http://code.google.com/p/universal-pos-tags/

The idea behind the definition of these universal tags [4] is that a set of (coarse) syntactic POS categories exist in similar forms across languages. These categories are often called universals to represent their cross-lingual nature [3] [10].
The tag set consists of the following twelve universal POS categories that exist in most languages:

NOUN (nouns),
VERB (verbs),
ADJ (adjectives),
ADV (adverbs),
PRON (pronouns),
DET (determiners and articles),
ADP (prepositions and postpositions),
NUM (numerals),
CONJ (conjunctions),
PRT (particles), '.' (punctuation marks)
X (a catch-all for other categories such as abbreviations or foreign words).

The authors claim that these parts of speech are the most frequent that exist in most languages, and have provided mappings to tag sets in 25 languages.

These three tag types constitute a representative set of de facto/best practise part of speech tag sets.
The difference between 1) and 2) on the one hand, and 3) on the other, is the level of descriptive granularity. The universal tag set delivers coarse-grained compatibility with tag sets covering many languages, whereas ISOCAT is linked to other available tag sets at a higher level of granularity though Relcat [3].
The resulting extended interoperability between their vocabularies and tag sets in the larger network provides a principled way towards re-use and comparison of part of speech information.

2. Representation in GATE output

The following annotation types represent the pos tag formats described above.

- **PTB tag**: the 'category' feature of annotation type 'Token'.

- **Universal tag**: the 'posUniversal' feature of annotation type 'Token'.

- Unambiguous **ISOCAT tag** (where a span only has one ISOCAT feature value; see figure 1 below):

  posIsocat1: the ISOCAT string identifier

  posIsocatPID1: the ISOCAT data category url.

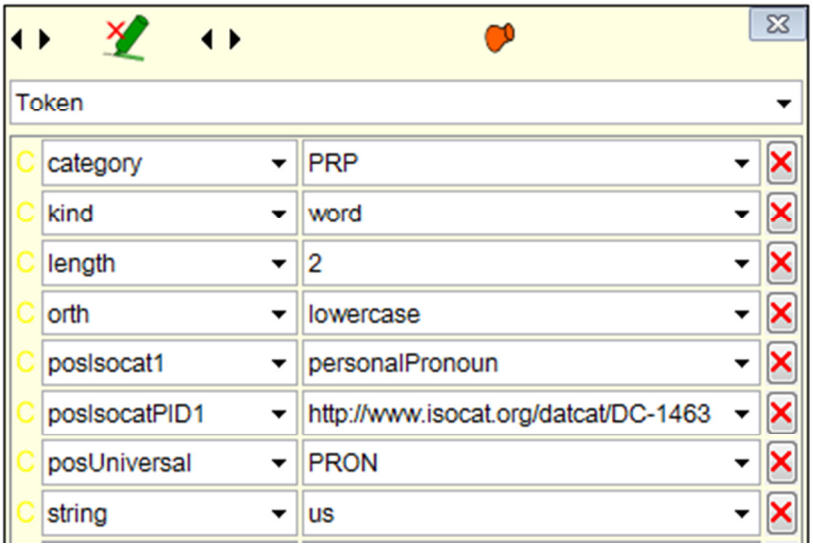

Figure 1: Unambiguous ISOCAT tag

- Ambiguous **ISOCAT tag** where a span has more thane one ISOCAT feature value. These come in two flavours:
    - AND (feature values are complementary, e.g. ‚verb‘ and ‘infinitive‘ in figure 2 below).

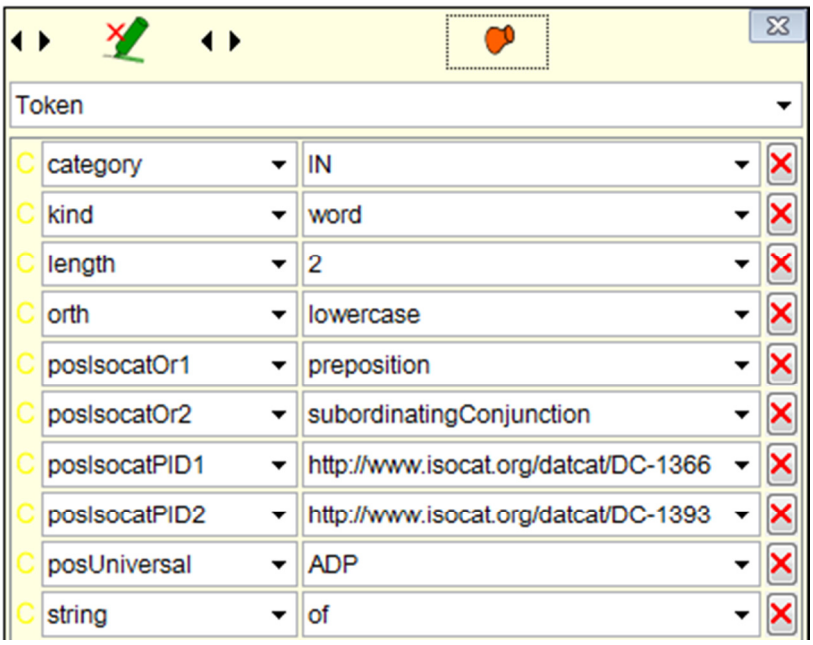

Figure 2: ISOCAT tag combination

- OR (feature values are mutually exclusive, e.g. ‘of‘ as a preposition of subordinating conjunction)

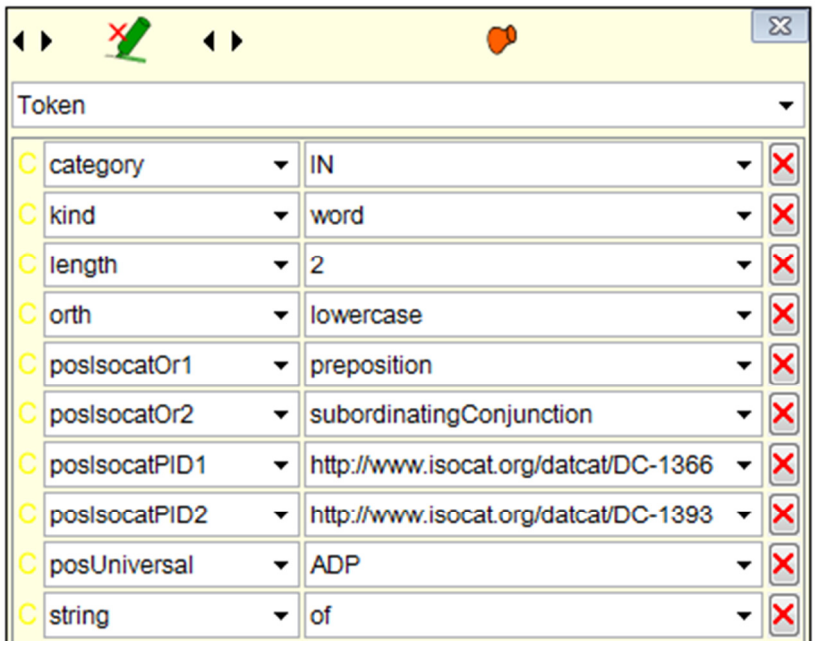

Figure 3: ISOCAT tag alternatives

The advantage of this representation is that the feature values are atomic and can be directly exported and referenced.

References

[1] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz: Building a Large Annotated Corpus of English: The Penn Treebank*, in* Computational Linguistics, Volume 19, Number 2 (June 1993), pp. 313—330

[2] Kemps-Snijders, M. Windhouwer, M.E. Wittenburg, P. Wright, S.E.: ISOcat: A Revised ISO TC 37 Data Category Registry. Presentation at the Conference on Terminology and Information Interoperability - Management of Knowledge and Content (TII 2008), Moscow, Russia (2008)

[3] I. Schuurman, I and Windhouwer, M.A., Explicit Semantics for Enriched Documents. What Do ISOcat, RELcat and SCHEMAcat Have To Offer? In: Proceedings of the 2nd Supporting Digital Humanities conference (SDH 2011). Copenhagen, Denmark, November 17-18 (2011)

[4] Slav Petrov, Dipanjan Das, Ryan T. McDonald: A Universal Part-of-Speech Tagset. LREC 2012: 2089-2096; http://arxiv.org/abs/1104.2086

[5] A. Carnie. 2002. Syntax: A Generative Introduction (Introducing Linguistics). Blackwell Publishing.

[6] F. J. Newmeyer. 2005. Possible and Probable Languages: A Generative Perspective on Linguistic Typology. Oxford University Press.

[7] Francopoulo, G., Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria: LMF for multilingual, specialized lexicons. In: LREC, Genova, Italy (2006)

[8] McCrae J., Spohr D., Cimiano P., Linking Lexical Resources and Ontologies on the Semantic Web with lemon. Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011), Heraklion, Crete (2011)

[9] Farrar, S. and Langendoen, T., A Linguistic Ontology for the Semantic Web. GLOT International 7 (3), pp 97 100 (2003)

[10] Peters, W., Establishing Interoperability between Linguistic and Terminological Ontologies, In: Oltramari, A.; Vossen, P.; Qin, L.; Hovy, E. (Eds.), New Trends of Research in Ontologies and Lexical Resources, Springer 2013.