A Comparative Evaluation of UNL Participant Relations using a Five-Language Parallel Corpus

Brian Murphy and Carl Vogel

Brian.Murphy@cs.tcd.ie*, Vogel@cs.tcd.ie
Department of Computer Science, University of Dublin, Trinity College

Abstract. In this paper we describe a manual case study in interlingual translation among five languages. Taking the UN Declaration of Human Rights in Chinese, English, German, Irish and Spanish, we annotated the five texts with a common interlingual logical form. We then studied four inventories of semantic roles (developed for both theoretical and NLP applications), including a subset of UNL's relations, and evaluated their suitability to describe the predicate-argument relationships found in the annotation. As a result, we make some suggestions for possible additions to the UNL relations, and propose that some of the existing relations be conflated or redefined.

1 Introduction

The work described here is part of a feasibility study on the use of semantic roles in interlingua-based machine translation. Our objective was to see if any set of semantic roles could give a description of verb-predicate relationships across a range of languages that would form an adequate basis for automatic generation.

The languages chosen were those that the authors have some working knowledge of (English, Chinese, German, Irish and Spanish), and include widespread and minority languages, both well and less-studied. The corpus used is the UN Declaration of Human Rights [1], a short text covering a broad range of topics in many languages (see Sect. 2).

From the literature on roles we selected four inventories (of which UNL's relations is one) that we considered to be well-enough developed for the annotation of unrestricted text. These inventories ([2,3,4,5] detailed in Sect. 4) were also chosen to be representative both theoretically and in terms of application to tasks such as machine translation and information retrieval.

After aligning the five language versions of the corpus, we manually annotated each article of the text with a language-neutral logical form (effectively a prototype interlingua) following the guidelines described in Sect. 3.1. The main part of the work then involved applying each of the role inventories in turn to the logical form and determining whether they satisfied three key criteria: coverage, differentiation and lack of ambiguity (Sect. 5). In other words, one should

* Supported by the TCD Senior Lecturer's Broad Curriculum Fellowship and Enterprise Ireland

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 64–76.

(1) a. Berufsschulunterricht müssen allgemein verfügbar gemacht werden ...
vocation-lesson must general available made to-be ...
'Professional education shall be made generally available ...' [Art. 26.1]

b. 父母对其子女所应受的教育的种类,有优先选择的权利
fùmǔ duì qí zǐnǚ suǒ yīng shòu de jiàoyù de zhǒnglèi, yǒu yōuxiān xuǎnzé de quánlì
parent to its children that should receive DE education DE type, has priority select DE right
'Parents have a prior right to choose the kind of education that shall be given to their children.' [Art. 26.3]

c. Is ionann na cearta atá acu ...
is same the rights that-are at-them ...
'They are entitled to equal rights' [Art. 16]

3 Interlingual Annotation

We manually aligned all 49 articles and sub-articles of the UN Declaration across the five languages, before adding English glosses (i.e. word-for-word translations, as seen in the previous examples) for all the non-English texts. The logical annotation of each article then proceeded on the basis of the English original and the four glosses, yielding over 500 predicates with almost 900 arguments. The aim was to arrive at a single, cross-linguistic logical form that, to the extent possible, adequately represented an article's meaning as expressed in all five versions. Although the result does not follow any of the five surface forms exactly, we aimed to abstract away from them only to the extent necessary to find a common representation.

To our knowledge there are no generally accepted guidelines for the manual annotation of unrestricted text with logical forms, as they are often theory or application specific. However, two sources proved useful. The Penn Propbank (a semantically annotated corpus) guidelines [7] have useful suggestions that we adopted for the treatment of phrasal verbs, support verbs and nominalizations. From cognitive science, Kintsch [8] gives an brief overview of annotation conventions for the 'microstructure' (roughly intra-sentence structure) of propositions, as used in comprehension modelling. We have broadly followed his treatment of negatives, modals, adjectives, adverbs and the status of propositions as arguments themselves.

3.1 Guidelines Developed

Negatives, modal verbs, adjectives and adverbs are expressed as one-place predicates with an event or object argument. As the focus of our studies is valency patterns, the quantification of objects was not annotated and noun phrases are rarely decomposed. Thus "all the rights and freedoms [Art. 2]" would be rendered as the atomic object *AllTheRightsAndFreedoms* as opposed to a form like [$\forall$x.[right(x) $\vee$ freedom(x)]]. Tense and aspect are not encoded.

Many of the conflicts between annotations suggested by individual language glosses are superficial, (e.g. near synonyms such as 'fair' (English) versus 'córa' (Irish: 'just') and 'equitativo' (Spanish: 'equitable')), in which case one of the lexicalisations is arbitrarily chosen. However, when there is a conflict in meaning we use two criteria to decide on a common predicate structure. Majority rule is one – for example in (2), the predicate 'presumed' won out, as it is used in both Spanish and English, and we judged it semantically close to 'regarded' (Chinese) and 'understand' (Irish), but significantly different from the German 'count'. Secondly, subject to majority rule, the most componential logical form available is used, as what is lexicalized in one language as a single verb may be a verb-argument complex in another. Hence, in the example below, the form *expel(U,Person,Country)* as suggested by the German version is preferred over *exile(U,Person)*.

(4) a. No one shall be subjected to ... exile [Art. 9]
b. 任何人不得加以...放逐
rènhé rén bùdé jiāyǐ ... fàngzhú
any person must-not be-made ... exile
c. Niemand darf ... des Landes verwiesen werden
no-one may ... the country expelled be
d. Ní déanfar ... aon duine ... a chur ar deoraíocht
not make ... single person ... that put in exile
e. Nadie podrá ser ... desterrado
no-one will-be-able to-be ... exiled
f. shall(not(expel(U,O1:Anyone,O2:Country))) belong(O1,O2)

We have not yet settled on semantic model of the formal language we use, but it resembles a higher-order logic, or a first-order logic with named Skolem functions.

4 Models of Semantic Roles

Semantic roles were first posited by linguists to describe the nature of meaning relationships among arguments and verbs in sentences. They correspond to a subset of UNL's relations. In this work we concentrate on so-called participant relations (see Table 1) as opposed to the more oblique circumstantial roles such as *manner, purpose* or *condition*, which are less commonly included in role inventories.

The earliest role inventories [9,10] were causally based and mirrored the grammar of argument structure quite closely (consider Fillmore's *agentive, dative,* and *objective* cases). Jackendoff went on to introduce a localist hypothesis [11] (or "thematic hypothesis") based on the extension of verbs (e.g. 'stay', 'go') and prepositions (e.g. 'from', 'to', 'at') of location and movement to more abstract situations (5). For example, information is viewed as *theme* ('story' in (5d) and by extension 'what' in (5g)) and holders can be viewed as *location* ('student' in (5d) and by extension 'document' in (5e) and 'mine' in (5f)).

by incorporating situation-specific roles such as *information* and *percept* (Table 2). She has made an extensive verb lexicon available [13], where each of the 11 thousand entries is annotated with argument syntax and role structure, using verb frames based on Levin's [14] semantic classes.

Table 2. Dorr's LCS roles

AG	agent	TH	theme
EXP	experiencer	INFO	information
SRC	source	GOAL	goal
PERC	perceived item	PRED	identificational predicate
LOC	locational predicate	POSS	possessional predicate
BEN	benefactive modifier	ISNTR	instrument modifier
PROP	event or state	PURP	purpose modifier or reason
MANNER	manner	TIME	time modifier

Sowa [4] has developed a model of roles for knowledge representation (see Table 3) based on Dick's [15] work in information retrieval, and Somers' Case Grid [16]. Sowa replaces the locational column labels (*Source, Path, Goal, Local*) of Somers and Dick with the four causes from Aristotle's *Metaphysics* (*Initiator, Resource, Goal, Essence*) and introduces six intuitive verb classes, which combined with several additional distinguishing features (such as animacy for differentiating *agent* and *effector*) correspond to more conventional roles.

Table 3. Sowa Roles

	Initiator	**Resource**	**Goal**	**Essence**
Action	Agent, Effector	Instrument	Result, Recipient	Patient, Theme
Process	Agent, Origin	Matter	Result, Recipient	Patient, Theme
Transfer	Agent, Origin	Instrument, Medium	Experiencer, Recipient	Theme
Spatial	Origin	Path	Destination	Location
Temporal	Start	Duration	Completion	PointInTime
Ambient	Origin	Instrument, Matter	Result	Theme

The model of relations used by UNL is more extensive, including logical operators such as AND and OR, and other novel roles such as BAS (*basis for expressing degree*) and SEQ (*sequence*). In particular it gives us a comprehensive treatment of the commitative roles CAG, COB and CAO (*co-agent*, *affected co-thing* and *co-*

INS instrument: instrument to carry out an event, e.g. "cut with scissors$_{INS}$"

The OBJ relation (i.e. *patient* role) is used for both clearly affected patients (e.g. the *Anyone* argument of *expel()* in (4f)) and for less affected participants such as the complements of psychological verbs (e.g. the *innocent()* argument of *presume()* in (2f)) and communication verbs (e.g. 'story' in (5d)). While this in itself may not be a problem, it may be missing significant syntactic generalisations. In several languages the tendency of a syntactic object to be promoted to a more prominent position, such as subject, seems in part determined by its affectedness. In the examples below the passive (8b) and 'ba'/'bei' (9b, c) variants of *enjoy(I,TheArts)*, all of which promote the object, are anomalous.[4]

(8) a. I$_{AGT}$ enjoy the arts$_{OBJ}$ [variation on Art. 27.1]
b. * the arts$_{OBJ}$ get enjoyed by me$_{AGT}$

(9) a. 我享受艺术
wǒ$_{AGT}$ xiǎngshòu yìshù$_{OBJ}$ [Chinese]
me enjoy art
b. * 艺术被我享受
* yìshù$_{OBJ}$ bèi wǒ$_{AGT}$ xiǎngshòu
art BEI me enjoy
c. * 我把艺术享受
* wǒ$_{AGT}$ bǎ yìshù$_{OBJ}$ xiǎngshòu
me BA art enjoy

Both [2] and [4] give a directional interpretation of these verbs, where the *enjoyer* above is a *goal* and 'the arts' a *source*. However examples from our corpus show that using the localist hypothesis (see Sect. 4) with these verbs does not generalise across languages. As we see below (10), in German our enjoyment is 'in' the arts, while in Irish almost the reverse is true – the enjoyment is 'at' us. As a result we suggest that a simple alternative is to use AOJ (roughly equivalent to *theme*) for non-affected syntactic objects. A more significant reworking would be to add the new roles of PRC (*percept*) and INF (*information*) following the practise of [3].

(10) a. Everyone$_{AGT}$ ... to enjoy the arts$_{OBJ}$ ... [Art. 27.1]
b. ... sich$_{AGT}$ an den Künsten$_{OBJ}$ zu erfreuen ... [German]
... self at the arts to enjoy ...
c. ... áineas na n-ealaíon$_{OBJ}$ a bheith aige$_{AGT}$ ... [Irish]
... pleasure of-the arts that be at-him ...
d. enjoy(Everyone,TheArts)

[4] A 'got' passive is used here as it cannot be mistaken for a non-passive adverbial sentence such as "he was unimpressed by the play". The star '*' indicates an idiosyncratic or ungrammatical form. The relation annotations shown follow UNL as it stands, rather than our proposals. BEI is an agentive marker and BA is an affectedness marker, both of which promote the object to a preverbal position.

event/entity distinction when it comes to locational roles, and the UNL relation PLC can be applied to both *Things* and *Events* (e.g. "a town$_{Thing}$ in Bavaria$_{PLC}$" and "She is$_{Event}$ in Bavaria$_{PLC}$"). In addition, it seems strange that the English prepositions 'from' and 'to' receive such special treatment, while the similarly common 'in' and 'of' do not.

Initially, the opposition of PLF/PLT for locations (e.g. (12) "return to his country$_{PLT}$") with SRC/GOL for states (e.g. (1a) "make education available$_{GOL}$") seems well justified.

(12) a. Everyone has the right ... to return to his country [Art. 13.2]
b. 人人有权...返回他的国家
rénrén yǒu quán ... fǎnhuí tā de guójiā
everyone has right return s/he DE country
c. Jeder hat das Recht ... in sein Land zurückzukehren
everyone has the right in his/her land to-return
d. Tá ag gach uile dhuine an ceart chun ... filleadh ar a thír féin
is at each every person the right to return to his country own
e. Toda persona tiene derecho ... a regresar a su país
every person has right to return to his/her country
f. entitled(O1:Everyone,return(O1,O2:Country)) belong(O1,O2)

However, some of the examples given in the documentation blur the distinction, in particular "go to Brussels$_{GOL}$" and "withdraw from the stove$_{SRC}$". It is not clear to us what basis there is for differentiating between 'his country' above as the final state of the entity 'Everyone' (GOL) or the final place of the event 'return' (PLT) – in both cases the ending of the event and the arrival of the agent happens in the same place at the same time. As a result, we suggest restricting SRC/GOL to non-spatial states only.

A more radical alternative would be to eliminate the SRC/PLF and GOL/PLT distinction altogether. We do not make a similar distinction for static locations (stative "famous in his field" and spatial "live here" both use PLC), and this is supported by [2,3] where spatial and stative end-points are conflated in *source/goal.*

5.3 Miscellaneous: POS, BEN

POS possessor: possessor of a thing, e.g. "the company's$_{POS}$ building"
BEN beneficiary: not directly related beneficiary or victim of an event or state, e.g. "be fortunate for you$_{BEN}$"

Possession is treated differently in UNL, depending on whether a genitive form ("that is my car$_{POS}$") or a possessional predicate ("I$_{AGT}$ have a pen$_{OBJ}$") is used. As with FRM/PLF and TO/PLT this seems like an unnecessary complication that none of the other inventories require. We also have to ask how agentive the subjects of verbs like 'have' and 'own' are – e.g. in what sense is the subject of "I have no money" an *agent*? Again we see that sentences of this type resist passivisation in English (13b) and the 'ba'/'bei' constructions in Chinese (14b, c). We suggest that possession be annotated as *possess(*POS,AOJ*)* following the practise of [3].

of BEN from adjuncts to also cover syntactic objects, and using POS for verbal as well as nominal structures that express possession. Finally we suggest several new situation specific roles (*recipient, effector, experiencer*) and explain how they might be of use in future versions of UNL.

References

1. United Nations General Assembly: Universal declaration of human rights. http://www.unhchr.ch/udhr/navigate/alpha.htm (1948) [Viewed December 2004].
2. Jackendoff, R.: Semantic Structures. MIT Press, Cambridge (1990)
3. Dorr, B.J.: Machine Translation: A View from the Lexicon. MIT Press, Cambridge (1993)
4. Sowa, J.F.: Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks/Cole, London (2000)
5. UNL Centre: The universal networking language specifications 3.2. http://www.undl.org/unlsys/unl/UNL Specifications.htm (2003) [Viewed December 2004].
6. UNL Centre: UNL manual. http://www.undl.org/unlsys/unlman/index.html (2001) [Viewed December 2004].
7. Kingsbury, P.: Propbank annotation guidelines. http://www.cis.upenn.edu/∼ ace/propbank-guidelines-feb02.pdf (2002) [Viewed December 2004].
8. Kintsch, W.: Comprehension: A Paradigm for Cognition. Cambridge University Press, Cambridge (1998)
9. Gruber, J.S.: Studies in Lexical Relations. Indiana University Linguistics Club, Bloomington (1965) Reprint of PhD Thesis.
10. Fillmore, C.J.: The case for case. In Bach, E., Harms, R.T., eds.: Universals in Linguistic Theory. Holt, Rinehart and Winston, New York (1968) 1–92
11. Jackendoff, R.: Semantic Interpretation in Generative Grammar. MIT Press, Cambridge (1972)
12. Saeed, J.I.: Semantics. Blackwell, Oxford (1997)
13. Dorr, B.J.: LCS database documentation. http://www.umiacs.umd.edu/∼ bonnie/LCS_Database_Documentation.html (2001) [Viewed December 2004].
14. Levin, B.: English Verb Classes and Alternations. University of Chicago Press, Chicago (1993)
15. Dick, J.: A Conceptual, Case-Relation Representation of Text for Intelligent Retrieval. PhD thesis, Department of Computer Science, University of Toronto (1991)
16. Somers, H.: Valency and Case in Computational Linguistics. Edinburgh University Press, Edinburgh (1987)