Towards a conceptual core for multicultural processing: A multilingual ontology based on the Swadesh list

Chu-Ren Huang, Laurent Prévot, I-Li Su and Jia-Fei Hong
{churen,prevot,isu,jiafei}@gate.sinica.edu.tw

Institute of Linguistics, Academia Sinica
Taipei, Taiwan

Abstract. The work presented here is situated in the broader project of creating of multilingual lexical resources with a focus on Asian languages. In the paper, we describe the design of the upper-level we are creating for our multi-lingual lexical resources. Among the current efforts devoted to this issue our work put the focus on (i) the language diversity aiming at massively multi-lingual resource, and (ii) the attention devoted to the ontological design of the upper level.

Keywords: ontology, lexical resource, multilinguality, Swadesh list

1 Introduction

The work presented is this paper is situated in the broader project of creating of multilingual lexical resources with a focus on Asian languages. When approaching the domain of lexical resources and their use in Natural Language Processing comes the question of the task repartition between the lexical and conceptual levels. This has been investigated for a long time under the light of philosophical or formal knowledge representation principles [5, 17]. Recently, the "ontological trend" generated an important amount of work concerning these ontologies from both knowledge engineering and computational linguistics perspectives. These projects often differ radically in the way they handle the ontology-lexicon interface but a common concern they share is the design of an upper level for the resource: EuroWordNet [20, 21], SIMPLE ontology [8], *Ω* (Omega) [16], OntoSem [12], SUMO-WN [14], OntoWordNet [4]. Among these projects, OntoSem is the first to have multilinguality explicitly on their agenda, while neither SUMO-WN nor Onto-WordNet has explicit design for multilinguality.

In this paper, we describe the design of the upper-level we are creating for our multi-lingual lexical resources. Our work put the focus on:

- the language diversity it is covering,
- the attention devoted to the ontological design of the upper level.

As for the language diversity, the multi-national project ("Developing International Standards of Language Resources for Semantic Web Applications") [19] in which the work take place regroups Japanese, Thai, Italian and Taiwanese teams for creating a multilingual resource aligned with the Princeton WordNet [3]. Moreover, within our team we benefit of the input from other languages such as Bangla, Malay, Taiwanese, Cantonese and Polish. There are two levels of development for these languages: for the languages represented by the project members (Japanese, Thai, Chinese and Italian) the goal is to create basic but significant core lexicons that can be compared to the Base Concept of EuroWordNet [20] and Global WordNet Grid[1]. About the other languages we only collected their respective Swadesh lists [18] in the perspective of building a minimal but massively multilingual lexical resource.

About the ontological aspect, in spite of some efforts devoted to the design of the upper level, the existing resources were not clear enough about their ontological commitment which was making them somehow difficult to compare. We aim here to compare them more thoroughly under the light of the recent works in formal ontology in order either to select the most appropriate model for the upper-level of multilingual lexical resources.

After this introduction, the next section will present the different methodologies that can be used for selecting a core lexicon while the section 4 describe the compilation of the Swadesh list for the languages considered. Then the section 3 describes our experiments for designing a prototype for the core upper level. We then investigate how the coverage of our resource can be extended (section 6).

2 Approaches for designing a core lexicon

Traditional approaches considered for establishing a compact list of basic terms (or *core lexicon*) can be divided into two categories according to their criteria for selecting the terms: *semantic primacy* and *frequency.* In addition in this section we propose a third way to be explored: the *universality* criterion.

2.1 Frequency criterion

The first intuitive idea for selecting a core lexicon is to uses statistical information such as word frequency. However, this naive approach of simply taking the most frequent words in a language is flawed in many ways. First, all frequency counts are corpus-based and hence inherit the bias of corpus sampling. For instance, since it is easier to sample written formal texts, words used predominantly in informal contexts are usually underrepresented. Second, frequency of content words is topic-dependent and may vary from corpus to corpus. Last, and most crucially, frequency of a word does not correlate to its conceptual necessity, which should be an important, if not only, criteria for core lexicon. The definition of

[1] See http://www.globalwordnet.org/gwa/gwa_grid.htm

a cross-lingual basic lexicon is even more complicated. The first issue involves determination of cross-lingual lexical equivalences. That is, how to determine that word `a` (and not `a'`) in language A really is word `b` in language B. The second issue involves the determination of what is a basic word in a multilingual context. In this case, not even the frequency offers an easy answer since lexical frequency may vary greatly among different languages. The third issue involves lexical gaps. That is, if there is a word that meets all criteria of being a basic word in language A, yet it does not exist in language D (though it may exist in languages B, and C). Is this word still qualified to be included in the multilingual basic lexicon?

A recent elaboration [22] proposed to use the notion of *distributional consistency* rather than crude frequency. This measure provides better result than other statistically based approaches but it requires balanced corpus of significant size to be applicable. Such corpora are only available for few languages and we would like to have a method that could be used with languages deprived from extensive resources.

2.2 Semantic primacy criterion

To answer about the lack of consideration of the "conceptual necessity" of the terms selected by the *frequency* approach, it is natural to consider more foundational work concerning knowledge organization. The idea of this approach is to determine a list of concepts that are semantic primitives (or atoms) that cannot be easily defined from other concepts. These concepts are located in the upper part of various hierarchical models. Each new distinction is made on the base of clear different semantic features. The main problem of these semantic primitives is their abstractness that make them rarely lexicalized (e.g 1stClassEntity in *EuroWordNet* top-level [20], NonAgentiveSocialObject in *Dolce* [11] or SelfConnectedObject in *SUMO* [13])[2]. These upper-levels will have a role to play in the design of our resource but they are not so useful for the constitution of the lexical core we are thinking of. We would like the basic building block of our resource to come from linguistic source, corresponding to the "linguist" ontology builder types described by Eduard Hovy in [6].

2.3 Swadesh list or the universality criterion

The lack of resources for most of languages led us to consider the Swadesh list [18] (reproduced as an appendix) as a potential core lexicon. The Swadesh list has been developed by Moriss Swadesh in the fifties for improving the results of quantitative historical linguistics. His list remains as a widely used vocabulary of basic terms. The items of the list are supposed to be as universal as possible but are not necessarily the most frequent. The list can be seen as a least common denominator of the vocabulary. It is therefore mainly constituted by terms that embody human direct experience. The list is 207 items long and is

[2] In this paper we use SmallCaps font for concepts and `TypeWriter` font for terms.

composed by the totality of the 200-item Swadesh first list, plus 7 terms coming from a 100-item list that Swadesh proposed later. This list is available for a great number of languages and its inclusion in the resources being collected in the context of the Rosetta project[3] warrants the quality and the maintenance of the resource. Moreover the Swadesh list items have been selected for their universality. Although quite different from the semantic primacy, this criterion ensure some kind of linguistic primacy that we are interested in.

These characteristics qualify the list has an interesting starting point for building a core lexicon in many different languages and for establishing easily the translation links. However, the methodology for establishing the list (essentially dictated by Swadesh's field work) introduces several issues that we have to deal with.

First, although made of lexical atoms, nothing prevents many other potential atoms to be discarded simply because of their lack of relevance for lexico-statistic purposes. This issue is specially important because it forbids us, when trying to propose a structure for the list, to think the Swadesh list as a definitive list of concept. As a consequence, a room for subjective appreciation remains open for introducing new concepts in the list.

The second issue results also from the initial purpose of the list. To be usable in field work context, the list concerns only direct human experience and avoids completely other foundational domains. On the other hand there is a richness for verbs describing human everyday activities that do not require modern tools.

Finally the Swadesh list, by its nature, has been established for spoken language in the context of face-to-face interaction.

3 Experiments: designing a core ontology from the Swadesh list

3.1 The experiments on Chinese

The Chinese Swadesh list was obtained by consulting with the Academia Sinica Chinese Wordnet group. One or more Chinese Wordnet entry for each item of the list were obtained, and non basic readings were eliminated. Subsequently, we obtain automatically the concept distribution of the items in SUMO taxonomy through SINICABOW,[4] a resource developed at the Academia Sinica which combines the Chinese wordnet, the Princeton WordNet and SUMO [13].

3.2 The experiments with English

About the English list we studied three different ways for building a taxonomy out of the simple list:

A. Keep the structure as minimal as possible by not adding any further (generalizing) concept in the list.

[3] See http://www.rosettaproject.org/ for more information.

[4] See [7] and *http://bow.sinica.edu.tw/* for more information.

B. Keep the structure as minimal a possible but also try to get a reasonable organization from a knowledge representation viewpoint.
C. Simply align the terms to SUMO ontology [13] and prune the result.

The first experiment (A) was not very conclusive since the list itself only includes very few words that are situated at different specificity level. The Swadesh items are indeed typically situated at the basic or generic level of categorization specificity [2]:p82. It is therefore expectable that they do not present a lot of taxomomic relations among them.

For the experiments (B) and (C), we proceeded in two steps:

1. Disambiguate the Swadesh List items by associating each of them to with one WordNet synset.
2. Create the taxonomy.

For (B), the taxonomy was created manually in a bottom-up fashion, by grouping the terms into more general categories while trying to keep the taxonomy as intuitive and minimal as possible. It resulted in about 220 classes organized in a preliminary taxonomy. Some generalization levels are missing since too few Swadesh items were corresponding to these areas.

In the case of the SUMO version (C), once we got the WordNet synsets the further mapping to SUMO was immediate once thanks to the mapping proposed in [14].

As for the technical aspects, our mapping operations have been done semi-automatically under Protégé[5] and more specifically with the help of ONTOLING[6] plug-in. The existing resources we used were WordNet 2.1 and an OWL transaltion of SUMO.[7]. The results of the experiences are available in OWL format on this website.[8]

4 Comparing lexicalization patterns

In addition of Chinese and English we compiled the Swadesh list for Bangla, Malay, Cantonese and Taiwanese from native speakers (students and colleagues). The universality aim of the Swadesh list was confirmed in this experiments. The Swadesh list is extremely well covered in the languages we studied. The only item that was said to not be lexicalized is `stab` in Bangla that is translated by `thiknagro-ostro-die-aghat-kora` which means literally *hit-with-a-sharp-instrument*.

An interesting issue has been raised by the Malay data in which several Swadesh items received the same Malay equivalent:

[5] For more information, visit *http://protege.stanford.edu/*

[6] For more information, see [15] and visit *http://ai-nlp.info.uniroma2.it/software/OntoLing/*

[7] Available at *http://www.ontologyportal.org/translations/SUMO.owl.txt*

[8] See *http://www.sinica.edu.tw/∼prevot/Swadesh/*

- `kaki`: both `foot` and `leg`
- `perut`: both `belly` and `gut`
- `hati`: both `heart` and `liver`
- `jalan`: both `road` and `walk`

For example `perut` corresponding to both `belly` and `gut` in the Swadesh list but for which the most natural translation is `stomach` can be compared to *nari-vuri* (`gut` again) in Bangla which is a compound word made of *nari* (small intestine) and *vuri* (large intestine).

These examples emphasizing the complexity of the lexical-conceptual relation. But might also raise the issue of cultural specificities at the conceptual level itself. In order to not loose such information our conceptual upper level cannot come only from experiments done in a given language or by a team of ontologists from a given culture. Such an language independent structure is likely to emerge in the upper level but we should not impose it as a starting point.

The structure presented in [21] also supports such a view. In Wordnets there are words and synsets (or word senses). The synsets across languages are not the same and their semantic organization differs. These synsets structures constitute real language-dependent lexical ontologies. It is worthwhile to consider these sense organizations coming from the languages before wrapping them up in an already established ontologies usually developed mono-culturally.

5 The problems encountered so far

5.1 Function words

A significant amount of words in the list (28 out of 207) are pronouns, demonstratives, quantifiers, connectives and prepositions. These words do not play a direct role in a taxonomy of the entities of the world. Unsurprisingly, many of them are absent from both WordNet and SUMO (e.g `you, this, who, and,...`). Quantifiers are present in WordNet (in adjectives) but placing them in the taxonomy is a thorny issue. SUMO-WordNet grouped them mysteriously under the EXISTS concept together with concepts such as `living`. About prepositions, some are present in WN (e.g `in`) but some other not (e.g `at`). In the beginning phase of the project, we simply decided to isolate all these words and to defer the discussion about them for later.

5.2 Ambiguities

The success of the Swadesh list is partly due to its under-specification and to the liberty it gives to compilers of the list. The absence of gloss results in genuine ambiguities, although some of them are partially removed through minimal comments added in the list (e.g `right` (correct), `earth` (soil)) and the implicit semantic grouping present in the list. More complex cases include terms like `snow` or `rain` that may refer to a meteorological phenomenon or to a substance.

In such cases we allowed ourselves to integrate both meanings in the taxonomy (e.g SnowSubstance *is-a* Substance, SnowFall *is-a* Phenomenon).

In this precise case, this ambiguity might be resolved by considering the semantic grouping that is sometimes proposed in the list (here `snow` appear together with `sky, wind, ice` and `smoke`).

For dealing with polysemy the solution could be to position the given polysemous *synsets* under several ontological concepts. However, placing in a taxonomy a term under two incompatible concepts results in an inconsistent resource. A way to deal with this problem, is to have only a few core meanings (ideally one) and derive the other senses from a richer relation network including other relations than hyperonymy. The generative lexicon [17] is an illustration of this possibility where the simple taxonomic link is replaced by four different relations.

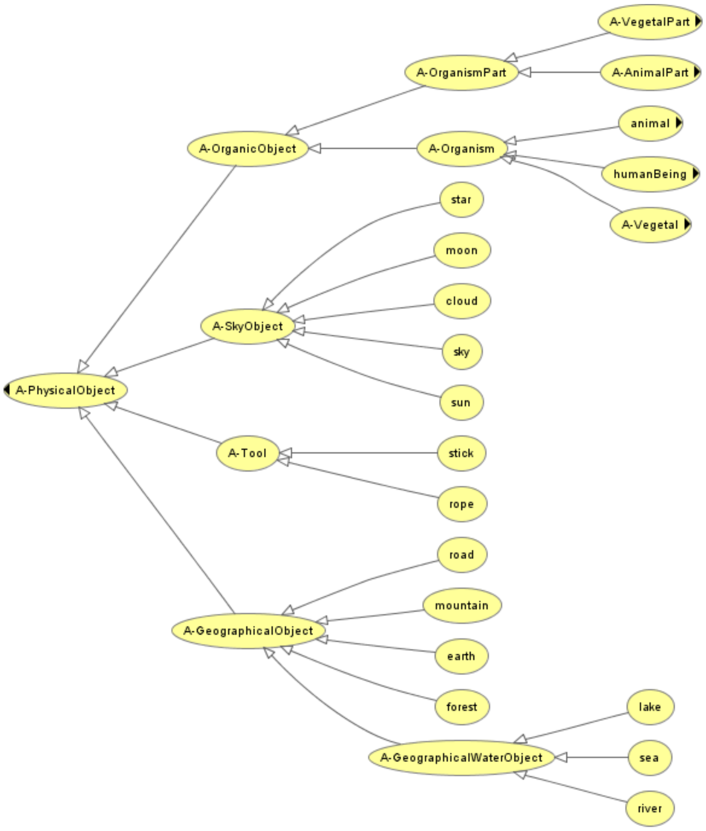

Fig. 1. Manual bottom-up taxonomy extrapolation (Method B), Physical object

5.3 Granularity heterogeneity and more general categories

Here the methodology chosen (A, B or C) introduces different issues. In the case of A, we actually did not succeed in identifying a structure where the nodes will be lexicalized by the items of the list. At best we get clusters than can be grouped under a general concept though extremely vague relation. For example, `sea, lake` and `river` might be grouped under WATER with option A. But so can be `rain, snow, ice` or `cloud` and why not having also `wet` and `drink`, `swim` or even `fish`. All these terms are *associated* with `water` but that do not qualify them for being equally positioned under WATER as a concept in a taxonomy. An on-going extension of WordNet concerns the addition of these loose links between the terms [1]. According to this study, such relations could remain unlabeled. However a step further could consist in identifying more precisely the nature of these "associations". For example, many of these terms refer to entities *constituted-by* WATER, others are *physical-state* of WATER or activities involving WATER. But adding these precisely links drive us away from the initial stage of our project.

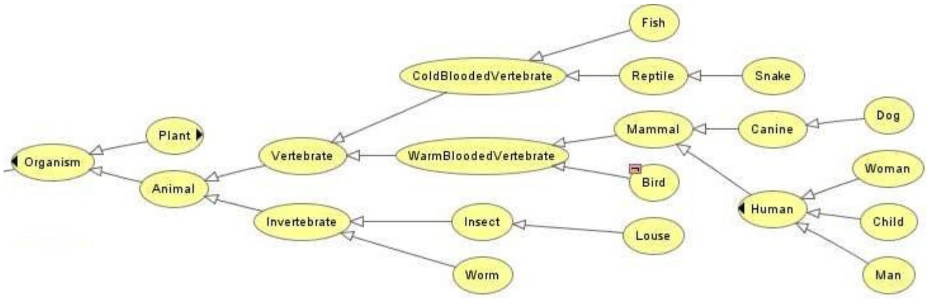

Fig. 2. Organic Object from method C

The options B and C actually takes us a step further away by introducing many new categories for disambiguating the terms and for accounting for intermediates levels such as BodyParts, Processes... (See Fig 1 and 2).

The ontologies resulting from (B) and (C) experiments, (B) has much flatter structure than (C). The ontology coming from the SUMO filtering includes actually a lot of intermediate levels that are not necessary for classifying satisfactorily the Swadesh items. As a consequence, the resulting structures need to be pruned and trimmed as illustrated in the figure 2.

5.4 Conceptual discrepancies?

The last issue concerns the discrepancies about the world conceptualization between on the one hand direct human experience viewpoint and on the other hand the modern scientific viewpoint. For example, which relations we will retain in our ontology for terms such as `sun, star` and `moon`. There is no such term as `satellite` in the list and nothing indicates that this concept is relevant for a direct human experience viewpoint. For now, while following the options B and C, we made the less committing choice. In this example we placed all the terms in question under the AstronomicalBody SUMO concept and under SkyObject for our own taxonomy proposal.

When turning to cross-cultural studies, it becomes clear there are different lexical (ar potentially conceptual) organizations for a given domain. See for example, the case of body parts in Rossel Island [9] or the one of geographical objects in Australia [10]. More examples (perhaps lexical only) are coming from the lexical gaps that are frequent as it has been noticed in the different lexical multilingual resources projects [20].

These issue highlight again the need to separate the lexical (*words*), semantic (*word senses* or *synsets*) and the ontological level (*formal concepts*). When facing a difference in lexical organization like the one observed in Malay. The word `perut` corresponds both to `belly` and `intestine`, there are four options for dealing with such a situation:

1. `perut` as a word having one "ambiguous" meaning ⟨*perut*⟩[9] corresponding to two different concepts in the ontology Belly and Intestine. Here a meronymy relation might holds between the two concepts –*Part-of*(Intestine,Belly)– which could explain the polysemy.
2. `perut` as a word having two meanings $\langle perut_1 \rangle$, $\langle perut_2 \rangle$, respectively mapped to Intestine and Belly. This will correspond to the SUMO-WordNet case since the sense division in WordNet is extremely fine grained.
3. `perut` as a word having one meaning ⟨*perut*⟩ requiring the creation of a new concept in the ontology corresponding roughly to the Belly and the Intestine together. In some case it might be difficult to define the new category on the base of other categories. Still we can assume as a working assumption that starting from a foundational ontology all new concepts are definable in terms of the ones already present. This model results in two types of categories in the ontology: stated (from axioms) and inferred (theorems). These apparently technical considerations are related to the discussions on the nature of the categories of the mental lexicon. These categories include both stable learned categories and other ones that need to be computed.
4. Finally, `perut` as a word having a vague meaning ⟨*perut*⟩ corresponding to an ontological concept with a weak characterization that might catch more word senses from different languages and that would not quite fit in the picture if we restrict too precisely the intented meaning of our vocabulary. Ontology is seen as a expression of shared ontological commitments as precise as possible for avoiding misunderstandings. Therefore, ontologists probably want to rule out this last option, keep these complexities at the "linguistic" level and start using ontology with the concepts and their definitions precisely established.

[9] In the rest of the paper the synsets will be written in italic within these brackets ⟨*synset*⟩.

Lexical gaps do not implies conceptual gaps and in many case it might be tempting to handle multilingual lexical discrepancies with complex mapping described (cases 1,2). However, we should not discard completely the case 3 and 4 specially when dealing with very different cultures.

6 Extending the word list

As emphasized earlier in this paper, the Swadesh list offer an interesting starting point for an universal core lexicon but is not enough by itself. In this section we compare the coverage of the Swadesh list with the one of the Base Concept Set [20, 21] as it is proposed by the Global WordNet Association[10]. Since both the Swadesh list and BCS are linked to SUMO, we are in position to compare the repartition of their mappings to SUMO.

The BCS synsets are mapped to 928 SUMO concepts, the most frequent mapping are presented in the table 1. In the table, the number of mappings from the Swadesh list is also provided. The number in parentheses corresponds to the number of indirect mappings (e.g `cut` is mapped to Cutting but Process is a parent of Cutting). In this example, although Process did not receive any mapping, many Swadesh items are classified under it. Below the double line of the table is included the SUMO concepts hosting as significant number (according to the list size) of Swadesh list mappings without being among the most common host for Base Concept synsets.

In both case the SubjectiveAssessmentAttribute is the most frequently mapped. In the case of the Swadesh list mapping we found all the adjectives that present a certain degree of subjectivity (e.g `bad, new, dirty...`) (See the documentation for this concept in figure 6). We expect that once a more comprehensive model for *qualities* (or properties) proposed in the ontology many adjectives should find a more satisfactory place in the model.

```
(documentation SubjectiveAssessmentAttribute "The &%Class of
&%NormativeAttributes which lack an objective criterion for their
attribution, i.e. the attribution of these &%Attributes varies from
subject to subject and even with respect to the same subject over
time. This &%Class is, generally speaking, only used when mapping
external knowledge sources to the SUMO.  If a term from such a
knowledge source seems to lack objective criteria for its attribution,
it is assigned to this &%Class.")
```

Fig. 3. Documentation for SubjectiveAssessmentAttribute SUMO concept

More significant are the SUMO categories for which the repartition of the BCS was strikingly different from the one of the Swadesh list. All categories

[10] See http://www.globalwordnet.org/gwa/gwa_base_concepts.htm

SUMO Concept	Mapping BCS	Mapping Swadesh
SubjectiveAssessmentAttribute	338	18
IntentionalProcess	93	1 (11)
Process	84	0 (64)
Motion	78	7 (25)
Device	70	0 (0)
Artifact	64	1 (2)
Communication	62	1 (2)
IntentionalPsychologicalProcess	51	0 (2)
RadiatingSound	46	0 (1)
BodyPart	42	16 (31)
Putting	41	0 (0)
Removing	40	1 (1)
Region	39	1 (9)
TimeInterval	38	2 (2)
Group	37	0 (0)
ShapeAttribute	36	4 (4)
Position	36	0 (0)
Text	35	0 (0)
TransportationDevice	34	0 (0)
Human	33	1 (4)
Increasing	32	1 (1)
part	32	0 (0)
SocialRole	32	0 (0)
Touching	20	4 (4)
Organ	10	8 (8)
Impelling	8	4 (4)
ColorAttribute	2	5 (5)

Table 1. SUMO concepts receiving the more mapping from BCS and Swadesh

related to Artifact, Device (and therefore TransportationDevice) are totally absent from Swadesh list but among the most frequent mappings for the BCS. In addition of this expected result, IntentionalPsychologicalProcess, Communication, Radiating Sound, Group, Position, SocialRole and Text are also almost absent from the Swadesh list. This suggest the direction in wich we sould extend the Swadesh list items in priority: social objects, mental objects, artifacts and communication.

Finally, some of the frequent Swadesh list item are not well represented for BCS. In the *ColorAttribute*the Swadesh list includes the five primary colors, BCS only includes $\langle shade\rangle$ and $\langle colored\rangle$. Another oddity is the the presence of $\langle wife\rangle$ but not $\langle husband\rangle$ in BCS while both are present in Swadesh list. These unexpected holes in the BCS shows that even a small list like the Swadesh might present some interesting suggestions for the design of the core lexicon.

7 Conclusion and future work

In this paper we investigate the idea of using the Swadesh list as a central resource for developing massively multilingual resources. Among the variety of on-going efforts on the development of multilingual resources, our project put the focuse on:

- Language diversity: the languages we consider are not only European languages as EuroWordNet [20] or SIMPLE [8].
- Solidity of the ontology: We do not want to commit our upper level too early to an existing ontology. We prefer to carefully compare the existing proposals, trying to understand the design choices for determining which ones are the most pertinent for the upper-level of a multilingual lexical resource.

As for the Swadesh list, we identified some limitations for this resource and emphasized some benefits of its usage. More precisely, it can be used as a good starting point for developing a linguistic ontology of direct human experience for a great number of languages. Such a resource is useful:

- (i) *per se*, for comparing different versions of the different lexical organizations (if there is more than one) and investigate the hypotheses of the relativist/universalist debate.
- (ii) as a first step for constituting a more applicative core lexicon for direct human experience that can should be integrated with core lexicon to form a full lexicon.

About (i), it is clear that more empirical experiments are needed in order to establish the structure underlying the list. An interesting approach could be to start with unlabeled semantic relations as described in [1] and later try to specify these relations according to their semantics.

About (ii), the Swadesh list, being limited to direct human experience and established in a spoken language context, has to be efficiently complemented by basic concepts of foundational knowledge areas (such as ARTEFACTS) for increasing its interest as a resource for NLP. Another important aspect is the integration of other relations than the taxonomic one in order to address the polysemy issue as described in 5.2. About this last point we are currently encoding the meronomic relations as well as various participation relations of objects to processes (thematic roles relations).

Finally for many languages other resources are simply not available. Developing such a micro-lexicon can be taken as a seed for developing a more significant lexicon around it. The idea is to have a simple upper level that is not disproportionated compare to the size of the lexicon but that can be seen as a first step toward the integration of a new languages in the Semantic Web. In this perspective the experiment B (manual extrapolation from a core lexicon) might be easier to perform than the integration of the terms in an already existing complex resource (experiment C).

To test further these ideas, we are currently working in collaboration with colleagues in Vietnam and the Philippines for extending the experiment to more languages.

Acknowledgment

We would like to thank the people involved in this project in Taipei, Kamrul Hasan, Sophia Lee, Siaw-Fong Chung, Tian-Jian Jiang, Katarzyna Horszowska Yong-Xiang Chen and the NEDO project members that provide the Swadesh lists. This research was carried out through financial support provided under the NEDO International Joint Research Grant Program (NEDO Grant).

References

1. Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. Adding dense, weighted connections to wordnet. In *Proceedings of the Third International WordNet Conference*, 2006.
2. W. Croft and A. Cruse. *Cognitive Linguistics*. CUP, 2004.
3. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
4. A. Gangemi, R. Navigli, and P. Velardi. The ontowordnet project: extension and axiomatisation of conceptual relations in wordnet. In *International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE)*, Catania, (Italy), 2003.
5. Jerry R. Hobbs, William Croft, Todd R. Davies, Douglas Edwards, and Kenneth I. Laws. Commonsense metaphysics and lexical semantics. *Computational Linguistics*, 13(3-4):241–250, 1987.
6. E. Hovy. Methodologies for the reliable construction of ontological knowledge. In *Proceedings of the 13th Annual International Conference on Conceptual Structures (ICCS 2005), Springer Lecture Notes in AI 3596*, 2005.
7. Chu-Ren. Huang, Ru-Yng. Chang, and Shiang-Bin. Lee. Sinica BOW (bilingual ontological wordnet): Integration of bilingual WordNet and SUMO. In *4th International Conference on Language Resources and Evaluation (LREC2004)*, Lisbon, 2004.
8. A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 2000.
9. S. C. Levinson. Parts of the body in yeli dnye, the papuan language of rossel island. *Language Sciences*, 28:221–240, 2006.
10. D. M. Mark and A. G. Turk. Landscape categories in yindjibarndi: Ontology, environment, and language. In *Proceedings of COSIT-2003, LNCS-2825*, 2003.
11. Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. Wonderweb deliverabled18, ontology library (final). Technical report, LOA-ISTC, CNR, 2003.
12. M. McShane, M. Zabludowski, S. Nirenburg, and S. Beale. Ontosem and simple: Two multi-lingual world views. In *ACL 2004 Workshop on Text Meaning and Interpretation*, Barcelona, Spain, 2004.

13. I. Niles and A. Pease. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine, 2001.
14. I. Niles and A. Pease. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03)*, Las Vegas, Nevada, 2003.
15. Maria Teresa Pazienza and Armando Stellato. The protégé ontoling plugin - linguistic enrichment of ontologies. In *Semantic Web 4th International Semantic Web Conference (ISWC-2005)*, 2005.
16. A. Philpot, E. Hovy, and P. Pantel. The omega ontology. In *Proceedings of ONTOLEX'2005*, 2005.
17. J. Pustejovsky. *The generative lexicon.* MIT Press, 1995.
18. Moriss Swadesh. Lexico-statistical dating of prehistoric ethnic contacts: With special reference to north american indians and eskimos. In *Proceedings of the American Philosophical Society*, volume 96, pages 452–463, 1952.
19. Tokunaga Takenobu, Virach Sornlertlamvanich, Thatsanee Charoenporn, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Chu-Ren Huang, Xia YingJu, Yu Hao, Laurent Prevot, and Shirai Kiyoaki. Infrastructure for standardization of asian language resources. In *Proceedings of ACL-COLING*, 2006.
20. P. Vossen, editor. *EuroWordNet: a multilingual database with lexical semantic networks.* Kluwer Academic Publishers, 1998.
21. P. Vossen. Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *International Journal of Lexicography*, 17/2, 2004.
22. Huarui Zhang, Chu-Ren Huang, and Shiwen Yu. Distributional consistency: A general method for defining a core lexicon. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, 2004.

Appendix: The Swadesh list

i thou he we you they this that here there who what where when how not all many some few other one two three four five big long wide thick heavy small short narrow thin woman man human child wife husband mother father animal fish bird dog louse snake worm tree forest stick fruit seed leaf root bark flower grass rope skin meat blood bone fat egg horn tail feather hair head ear eye nose mouth tooth tongue fingernail foot leg knee hand wing belly guts neck back breast heart liver drink eat bite suck spit vomit blow breathe laugh see hear know think smell fear sleep live die kill fight hunt hit cut split stab scratch dig swim fly walk come lie sit stand turn fall give hold squeeze rub wash wipe pull push throw tie sew count say sing play float flow freeze swell sun moon star water rain river lake sea salt stone sand dust earth cloud fog sky wind snow ice smoke fire ashes burn road mountain red green yellow white black night day year warm cold full new old good bad rotten dirty straight round sharp dull smooth wet dry correct near far right left at in with and if because name