# Using the Swadesh list for creating a simple common taxonomy

Laurent Prévot, Huang Chu-Ren, Su I-Li

Institute of Linguistics, Academia Sinica
Nankang, Taipei, Taiwan 115
{prevot,churen,isu}@gate.sinica.edu.tw

**Abstract.** The long-term goal of the research described in this paper is to develop a multilingual language resource linked to the Princeton WordNet. The paper describes the experiments we are conducing for determining a basic vocabulary and for designing a language-independent core for the future resource. More precisely, in this paper we use the universality of the Swadesh list [15] for selecting it as a basic core vocabulary and we present several options for designing a minimal upper ontology underlying the list.

**Keywords:** Language Resource, Lexicon, Multilinguality, Ontology, Interlingual index, Swadesh list

## 1 Introduction

One of the main goal of 'Developing International Standards of Language Resources for Semantic Web Applications' [16], an international project sponsored by Japan's NEDO foundation, is to implement standards of language resources that can be very robust when applied to different languages of the world. In addition, the project concentrates on Asian language resources. Hence the project plans to construct lexica of Japanese, Mandarin, Thai and Italian; and integrate them as parts of a multilingual resource linked to the original Princeton WordNet (WN) [2]. It is therefore comparable with projects such as EuroWordNet [17], in which the coherence and homogeneity across different languages remained the main issue that hampered the true inter-operability of the final resource. Since the project is exposed to these risks by its design, a priority for the project members is therefore to tackle this issue from the very beginning. For this purpose, one of the measures taken is to build jointly an upper-level ontology[1] that will play the role of a structured interlingual index. The NEDO participants are currently exploring several ways for selecting a basic vocabulary that will serve as a starting point for designing this language-independent core of the resource. This paper describes some of the preliminary experiments we

[1] In this paper, ontology has to be understood in a light-weight sense. At this development stage the 'ontology' is much more a simple taxonomy than a complete axiomatic theory.

are currently conducing. More precisely we are (i) using the Swadesh list [15] as a basic core vocabulary and (ii) exploring the possibilities offered by this list for creating a simple common ontology. These practical objectives confront us however with a complex discussion of contemporary linguistics, namely the relativist/universalist debate [4], and more precisely its consequences for the lexical organization. In the context of this project the existence and the nature of a common "universal" structure is a background question that we would like to contribute to answer.

## 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*.

### 2.1 Semantic primacy criterion

This approach proposes to constitute a list of term that are semantic primitives (or atoms) and cannot therefore be easily defined by using other terms. Under this approach, the terms appearing frequently in definitions gloss are good candidate for being integrated in the core lexicon. The main problem of this approach is that the upper level of existing ontologies are generally fairly abstract concepts that are rarely lexicalized 1stClassEntity in EuroWordNet or SelfConnectedObject in SUMO[10], and that are intuitively far from being member of a *core* lexicon.

### 2.2 Frequency criterion

The second approach uses more statistical data such as word frequency. However, simple word frequency is not good criterion for selecting the basic terms. A recent elaboration [18] proposed to use the notion of *distributional consistency*. This measure provides better result than other statistically based approaches but it requires balanced corpus of significant size. 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.

### 2.3 Swadesh list or the universality criterion

The lack of resources for most of languages led us to consider the Swadesh list [15] (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 attempt has not been very conclusive but the 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 composed by the integrality 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[2] 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:

- Although made of lexical atoms, nothing prevents many other terms to atoms too but 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 posit a stable 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 the set of cultural context Swadesh worked, 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 in a non-industrialized setting. This point, is however not worrying since the domain of the list is somewhat well-defined and it will be easy to figure out in which direction the list has to be extended for getting closer to a more usable resource for natural language procesing.
- Finally the Swadesh list, by its nature, has been established for spoken language in the context of face-to-face interaction.

## 3 The experiments

### 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,[3] a resource developed at the Academia Sinica which combines the Chinese wordnet, the Princeton WordNet and SUMO.

[2] See http://www.rosettaproject.org/ for more information.
[3] See [6] and *http://bow.sinica.edu.tw/* for more information.

### 3.2 The experiments on English

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

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

The options (B) and (C) were performed in two steps: (i) disambiguate the words of the list by mapping them to WordNet synsets and (ii) create the taxonomy either manually (B) or automatically (C). In case of (C) the further mapping to SUMO ontology is immediate thanks to the mapping proposed in [11]. In case of (B) after disambiguating the terms, we proceed, 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.

Our mapping operations have mostly been done semi-automatically under Protégé[4] and more specifically with the help of ONTOLING[5] plug-in. The existing resources we used were WordNet and the Protégé translation of SUMO.[6] The results of the experiences are available in OWL format on this website.[7]

## 4 The problems encountered so far

### 4.1 Function words

A significant amount of word of 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. The 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.

### 4.2 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

[4] For more information, visit *http://protege.stanford.edu/*
[5] For more information, see [13] and visit *http://ai-nlp.info.uniroma2.it/software/OntoLing/*
[6] Available at *http://ontology.teknowledge.com/*
[7] See *http://www.sinica.edu.tw/∼prevot/*

ambiguities, although some of them are partially removed through minimal comments added in the list (e.g *right (correct), earth (soil)*) and the implicit preliminary 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 (in this precise case, `Snow` appear together with `sky, wind, ice` and `smoke`). However, more generally speaking, current lexicographic works argued convincingly that they are not such things as word senses in a traditional sense. These senses are convenient lexicographic artifacts that do not resist to deep empirical studies [7]. An alternative way to deal with *"word senses"* proposed by Kilgariff (but that can be identified also in the FrameNet project) is to consider a sense for a word as a set of occurrences in similar contexts. However such an approach requires annotated corpora of significant size. Our approach deals with languages for which resources are not available.

An idea for representing this polysemy without multiplying the nodes of our resources consist in attaching the polysemous term under several ontological concepts. However as explained in [3], placing in a taxonomy a term under two incompatible concepts results immediately in an inconsistent resource. A way to work around this problem, is to create underspecified polysemous nodes that are related to other categories with other relations than the taxonomic "is-a". The generative lexicon [14] is an illustration of this possibility where the simple taxonomic link is replaced by four different relations (not all allowing for inheritance along the relations).

### 4.3 Granularity heterogeneity and more general categories

Another class of problems we encounter was the granularity heterogeneity of the list. For example `dog` and `animal` are included but they are no species proposed for `bird, fish, tree`. General categories have been avoided for the initial purpose of the list. We have now to handle this heterogeneity when dealing with taxonomic issues.

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` 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 unlabelled. However we consider that 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 typically `water`. But adding these precisely links drive us away from our core vocabulary.

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.... These concepts seem necessary for providing some order to the list of words but there are not the basic lexical terms we were thinking of. These more abstract entities, reveal there existence not directly in the lexicon since people do not need to refer to these general entities but more deeply in the language through grammatical constraints for example. Although natural language semantics has shown that semantic categories play a crucial role in all the level of interpretation. These categories, often delineated by syntactic and semantic behavior, do not have to correspond to lexical entries. There is no contradiction for an ontology to be linguistically biased while retaining concepts as nodes and allowing for complex mappings between ontological concepts and lexical items.

### 4.4 Conceptual discrepancies?

The last issue concerns the discrepancies about the world conceptualization between on one hand direct human experience viewpoint and on the other hand the modern science 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.

Moreover, it is clear that it exists different lexical organization for a given domain. See for example, the division studies of bodyparts in [8] or the one of geographical object in [9]. Closer to our experiment, EuroWordNet team noticed that in Dutch there is no concept for a generic `container` while in other languages this term was available for being included in the core lexical structure. Finally, see [12] for a discussion of the [5] example of `wood, tree` and `forest` in French, German and English. As a conclusion on this topic, the nature of the list and structure we are aiming too is essentially linguistic and do not pretend to say anything about deeper cognitive structure.

This issue highlights again our need to separate the lexical and the ontological level. The fact that a language selected a term does not mean that concepts that did not received similar label are absent. Moreover the permeability between word senses [7] and the robustness of the semantic system allows for an adaptation of the language. In the `'sun'` example, there are good chances that a language using `sun` to refer to the instance only will adapt it easily for more scientific usages.

## 5 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. We identified some limitations for this resource and emphasized the benefits of its usage. More precisely, the Swadesh list can be used as a 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.

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 domains (such as *artefacts*) for increasing its interest as a resource for NLP. Another important aspect is the integration of other relations than the taxonomic "is-a" in order to tackle the polysemy issue as described in 4.2

Finally, we are currently working on the alignment of our list with the ones made in other sites (Japan, Thailand, Italy) and other languages (including Cantonese, Bengali and Malaysian).

## 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. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
3. Nicola Guarino. Some ontological principles for designing upper level lexical resources. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada, editors, *Proceedings of First International Conference on Language Resources and Evaluation*, pages 527–534. ELRA, 1998.
4. J. J. Gumperz and S. C. Levinson. *Rethinking Linguistic Relativity*. Studies in the social and Cultural Foundations of Language. Cambridge University Press., 1996.
5. L. Hjelmslev. Dans quelle mesure les significations des mots peuvent-elles être considérées comme formant une structure? In *Proceedings of the 8th International Congress of Linguists*, 1958.
6. 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.

7. Adam Kilgarriff. "i don't believe in word senses". *Computers and the Humanities*, 31:91–113, 1997.
8. S. C. Levinson. Parts of the body in yeli dnye, the papuan language of rossel island. *Language Sciences*, 28:221–240, 2006.
9. D. M. Mark and A. G. Turk. Landscape categories in yindjibarndi: Ontology, environment, and language. In *Proceedings of COSIT-2003, LNCS-2825*, 2003.
10. 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.
11. 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.
12. S. Nirenburg and V. Raskin. Ontological semantics, formal ontology and ambiguity. In *Proceedings of FOIS 2001*, 2001.
13. 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.
14. J. Pustejovsky. *The generative lexicon.* MIT Press, 1995.
15. 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.
16. 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.
17. P. Vossen, editor. *EuroWordNet: a multilingual database with lexical semantic networks.* Kluwer Academic Publishers, 1998.
18. 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