Polysemy and Categorization: 
Implications to Corpus Documentation*

Lily I-wen Su and Shuping Huang 
National Taiwan University

Studies of the endangered languages depend heavily upon reliable fieldworks and linguistic documentation. A common method used in meta-linguistic glossing is to ask the informant to translate the object language into the meta-language. This strategy, as warned by Matthewson (2004), can be problematic, and the present study will focus on two of the problems: (a) the confusion of semantic meaning and pragmatic implicature, and (b) the mismatch of sense classification in object language and meta-language. Based on the theory of categorization, in particular Principled Polysemy (Tyler and Evans 2001), we hold that pragmatic implicatures, triggered by textual/situational/cultural contexts, should be represented in free translation, and morph-by-morph glossing should be saved for semantic cores. Meta-linguistic glossing often contains cultural-specific information. When information of this nature is not well-documented, corpus thus established fails to capture information indispensable for the understanding of the language.

Key words: polysemy, categorization, comparative semantics, contextualization, corpus development

1. Introduction

The attempt to describe a particular sense often requires people to face the challenge often associated with the account of semantic extension — whether those senses treated as polysemous ‘extensions’ from a core are necessarily best accounted for as such. To begin with, there are often mismatches of conceptual categorization in different languages, a problem typically overlooked by inter-language translation. For example, the English phrases put on and put in constitute a large spatial space of placing one thing on top of the other. However, this space is partitioned differently in Korean into five categories denoted by nohta, ssuta, pwuchita, kkita, and nehta. (Bowerman and Choi 2003). To gloss these Korean lexical items either as ‘put on’ or ‘put in’ cannot satisfactorily reflect the spatial configurations structured by those lexical items.

* This paper is dedicated to Professor Paul Jen-kuei Li on his seventieth birthday.
The implication of this challenge on studies of the endangered languages can not be undermined, especially when it comes to the corpus studies of such languages. The success of a research on endangered languages based on corpus data cannot be guaranteed unless reliable fieldwork, especially in terms of its methodology and documentation, is secured. Since issues related to the treatment of the lexical items start early in the beginning of a corpus study, a key question here is to answer the fundamental question of ‘How should the meaning of a lexical item be determined?’ (Samarin 1967). In spite of the fact that Matthewson (2004) warns that linguists should be careful about the use of translation, one intuitive way to handle the problem is to ask the informants for meaning translation via the use of a meta-language known both by the speaker and the informant/consultant. The present study aims to highlight some empirical concerns of semantic glossing, with methodological implications on language documentation and corpus development.

Informants often provide context-situated interpretation for each occurrence of the same lexical item, yielding polysemous readings of a lexical item, which is counter-intuitive to what the native speakers of the language would have conceptualized. When Riemer (2001) claims that it is theoretically not probable to enlist a new sense for a lexicon every time when it is used in a novel way, he is in fact pointing out the very problem described here. Language is often a reliable way to reflect speaker’s conceptual categorization. Objects, events, states, and even grammatical categories that are named by the same linguistic form may share some ‘similarities’ or ‘inferential links’ that the speakers attempt to emphasize — they are in the same category (Lakoff 1987). How semantic glossing can reflect the links between those contextually-activated interpretations is a great challenge to corpus development.

Furthermore, the meaning of a lexical item is often triggered by information that is culturally specific. For example, the lexical item kitten in English creates a stereotypical image of being sexually attractive when it is used to refer to a woman — an image with a positive connotation. Such is however not the case with Arabic and Persian societies — these cultures impose negative connotation for such association (Talebinejad and Dastjerdi 2005). Though linguists typically disagree with Malinowski’s (1923) approach to incorporate ethnography in corpus documentation, the understanding of semantics is inevitably interwoven with culture, speech situation, and linguistic contexts. While word-by-word glossing may fail to indicate cultural values and considerations, corpus users may fail, as a result, to correctly grasp and interpret its meaning, by itself or used in its proper linguistic context. How to facilitate such use into corpus is another significant issue that corpus developers should take into account.

In the following sections, we will cite cases from our NTU Spoken Corpus of Formosan Languages, with the purpose to illustrate how problems of this nature create a
real challenge to semantic glossing. We would claim that the study of semantics should go hand in hand with pragmatics so that considerations involving situational, textual, and cultural factors may shed light on corpus glossing. Contexts have great impact and bearing on lexical meanings, which is an aspect that one should take into consideration if one is to maintain a unifying semantic coding in a corpus.

2. Polysemy and categorization

Concepts encoded in natural language can be vague, but this is not necessarily true of their semantic description. Cognitive linguistics is not only interested in what constitutes the center of a category, but also in how this center can be extended to peripheral cases, that is, how far the extension can go.

Of particular interest to the study of concept is categorization. Categorization is part of our experience, and inclination to categorize is a feature of the way that we interact in our daily lives with the objects that surround us. Much of human abstract thinking, or cognition in general, is involved in the process of categorizing. Be it objects, concepts, etc., we seem to make sense of the world by creating little boxes in which to place various divisions of it.

Early in Aristotle’s age, it has been found that there is a systematic way in which people categorize entities or events into categories. Carolus Linnaeus further established a scientific taxonomy of living things, which represents the hierarchical structure of categorization (cited from Ungerer and Schmidt 1996). In traditional views, scientific taxonomy is considered an absolute way of categorization, and all members of a category are of equal status.

Revolutionary efforts have been made by Eleanor Rosch to show that folk categorization can differ greatly from a scientific one. In her experiments, some members of a category stand, due to conceptual saliency, as ‘good exemplars’ which often function as the reference point when compared to other category members. For instance, birds like a robin are more ‘birdly’ in English than a penguin or an ostrich (Rosch 1973, 1978, 1983). Such membership asymmetricity later assumes the status of the so-called ‘prototype effect’ in her works.

Notions of categorization have been found to be highly relevant to studies of language structures (Taylor 1995). Labov (1973:342) explicitly states that ‘If linguistics can be said to be any one thing, it is the study of categories: that is, the study of how language translates meaning into sound through categorization of reality into discrete units and sets of units.’ The result of categorization is mental lexicon (Ungerer and Schmid 1996), which reflects clustering of meanings by the same label. Concepts encoded in the same or related linguistic forms represent language users’ conceptualizations of
A linguistic unit can thus be considered as a category which groups together a variety of disparate members and establishes at the same time equivalence between these members — the equivalence lies in their all being deemed to be members of that category (Ellis 1993).

Based on the psychological assumptions of categorization, Lakoff (1987) holds that human beings organize knowledge into a mental structure, called an Idealized Cognitive Model (ICM), with the center of an ICM constituted by the idealized prototype of a given category. Some senses of a lexical item are more central (or prototypical), whereas others are radial. For example, bank where money is deposited is more central than blood bank in terms of the category led by the lexical item bank. Members of a lexical category have chaining relations (Austin 1961), or family resemblance (Wittgenstein 1963), linked by metaphor, metonymy, propositional, and image-schema (Lakoff 1987). The term ‘family resemblance’ denotes that there need not be commonly shared characteristics for all members, but any two of them are chained to each other via cognitively plausible links. An English example arm is extracted from Blank (2003). They are re-arranged and analyzed here.

(1) a. My arms ache from carrying this bag. (prototype)
   b. the arm of a jacket (metonymic; image-schematic)
   c. the arm of a chair (image-schematic)
   d. an arm of water (image-schematic)
   e. an arm of an organization (metaphor)

The prototypical, central meaning of arm refers to the upper limbs of the human body, whereas other meanings are radial meanings not easily retrievable without relevant contexts. The difficulty increases from (1b-e), because the interpretation with the prototypical use (1a) becomes less and less transparent. The different meanings shown by (1a-e) may not share necessary and sufficient similarities, but they are chained to each other as a ‘family’ through different kinds of cognitive mechanisms indicated inside the parentheses. (1) is a case of lexical polysemy, adopting Taylor’s (1995:99) definition: polysemy is an ‘association of two or more related senses with a single linguistic form.’

Tyler and Evans’ (2001) ‘Principled Polysemy’ (PP) assumes a somewhat similar center-radial belief, although it is the notion of ‘protoscene,’ instead of the so-called ‘prototype,’ that is proposed to represent a conceptual category. Protoscene stands for a centrality abstracted from the various uses denoted by one single linguistic form, which can be described in a systematic framework when Langacker’s Cognitive Grammar (1991, 1999) is adopted. PP suggests that by changing the profiles or perspectives of a
‘protoscene,’ language users may enrich or impoverish the semantic elements to derive new meanings due to high frequency of use. Tyler and Evans’ idea has had a great impact on semantic studies: it provides a way to integrate diachronic semantics with its synchronic counterpart, although PP is actually intended for explanation of synchronic polysemy. From a diachronic perspective, meanings emerging from the contexts are possible candidates of a lexical item’s coded meanings. The new coded meaning co-exists with the ‘core’ ones, which results in the so-called ‘layering’ (Hopper 1991) of senses, where no observable link can be spotted at the first sight. The two senses of the English word *bar* — ‘piece of long-shape solid material’ and ‘barrier in the lawcourt separating the judge, prisoner, lawyers, etc.’ — serve as a good example for this point. In a synchronic sense, changes of construal can well lead to different senses (semantic coded meaning) and forces (pragmatic implicatures), a distinction maintained by Löbner (2002).

3. Linguistic categorization and the protoscene

In this section, we will show how various meanings represented by the same linguistic form may constitute a single semantic category. We will illustrate the point by a Saisiyat example. During our elicitation of Saisiyat, we have often been puzzled by *nahan*, a lexical item whose meaning is rather confusing. When we ask the informants for its meaning, we are given for each occurrence different meaning based on interpretation according to context-situated appropriateness. Based on the informants’ direct translation, five different translated equivalents are found: ‘again,’ ‘another,’ ‘still,’ ‘first,’ and ‘a while.’ The direct translation method renders different glosses for this very lexical item. This situation seems to suggest that the various uses of *nahan*, as indicated by their English glosses, are not related: their sharing the same linguistic form may be regarded as a sheer accident because what we have is a case of homophone. This view is quite unlikely if we accept the notion of linguistic categorization by Taylor (1995) and Lakoff (1987): the chance that many unrelated words may share the same linguistic form is slim. Therefore, our first question is: Are those uses related, and if so, how?

Another important question that we would address in the paper is: How necessary is the distinction between ‘sense’ (semantic coded meaning) and ‘force’ (pragmatic implicature)? The word ‘sense’ refers to the ‘discrete sense’ in Cruse (2000), but ‘force’ may be facets, subsenses, perspective, local senses, or even contextual modulations in Cruse’s terminology, a feature not discrete enough to be assigned an independent lexical meaning. That is, forces are not stored meanings in our mental lexicon — they are modulation of a construal existing ephemerally in the context. Put it in Riemer’s (2001) words, we may simply deem it common for a lexical item to be used in a novel way to
achieve some sort of effects, whereas it is economically unpractical to enlist each of the novel use as a new sense.

One way to test if the meanings denoted by *nahan* are cognitively and/or interactionally activated is to verify their relation and to determine its sense and force by appealing to Tyler and Evans’ Principled Polysemy (Tyler and Evans 2001). We would first illustrate the construal of each use to see if all five uses (represented by the five translated terms) share cognitive unity.

Example (2) is taken from a Saisiyat legend, a story that says one’s ancestors never died — the Saisiyat people believe their ancestors would molt and become rejuvenated again when decrepit. The word *nahan* in IU 13 receives the interpretation ‘again’ because a discontinuous phase exists between the men’s youth — one before and one after molting.

(2) molaw
12 ...(1.0) m-olaw ila
AF-molt PFV
13 ... so: m-olaw kita’en ma’ ‘al’alak ila *nahan*
if AF-molt see-PF also young PFV **again**
‘(Once the men grew old), they molted and looked young again.’

Following the conventions of Langacker’s Cognitive Linguistics (1991, 1999), the viewing arrangement of (2) can be sketched as Figure 1. The construal includes a trajectory (TR) in sequential relation to the landmark (LM), and the icon for TR is in the same shape and color as LM to show that it is the repetition or reduplicate of LM. The scope of predication (dashed frame) is limited to TR, but LM is hinted in the maximum scope of predication (solid frame).

![Figure 1: The viewing arrangement of *nahan* ‘again’](image)

In example (3), the speaker is talking about the breeds of deer he has kept. The speaker provides a listing, and he recalls another kind of similar deer though he forgot
the name after mentioning several members of deer breeds. The context where the gloss ‘another’ is used shares the same contrual (Figure 1) as ‘again’ in (2).

(3) Life
59. …(1.6) o: ‘aehae’ nahan sinraehoe isaa yao ma’
   DM one another species that lsg.nom too
60. .. aw ‘oka’ ila sekela atomalan
   FIL NEG PFV AF.know really
   isaa ‘aehae’ nahan ka== sinraehoe
   there one another NOM species
   ‘Another kind of deer I’m really not familiar with.’

(2) and (3) mainly differ in their predicate types. The ‘another’ reading is specific to numeric predicates, indicating shift from a temporal axis to a nominal one. Such shift of axis echoes the distinction Langacker (1991) makes: ‘summary scanning’ vs. ‘sequential scanning.’ Sequential scanning profiles the scene as comprising of the component activities in series, a perspective that better explains the grammatical category commonly known as ‘verb.’ Summary scanning is on the other hand a snapshot of the entire event, usually on the final ending state, which better explains the grammatical category known as ‘nouns.’ The protoscene represents a way of viewing the different profiled scene expressed by (2) and (3).

Example (4) is another Saisiyat legend talking about a goddess, who was married to a Saisiyat young man and now wanted to visit her parents’ home. The gloss ‘still’ provided by our informant seems to indicate the informant’s interpretation of the goddess’ longing for home-going as an incessant one.

(4) Kathethel 1
73. ...(0.9) isaa rwasek ila== isaa hN
   that AF.live PFV that FIL
74. ...(1.2) koza ila kin honahnge hiza ma’
   what PFV very long.time that also
   ‘It has been a long time after they got married.’
75. ... ka==
    FIL
76. ...(1.1) hiza minkoringan ‘am==
    that woman FUT
77. .. lobih k<om>osa ‘am rima’ kala yaba nahan
    AF.return <AF>say FUT AF.go LOC father still
78. ... nisia
   3SG.GEN
79. ... rima’ ray ‘inkahoe lan
   AF.go LOC married.woman’s.parents’.home
   ‘After getting married for a long time, the woman asked to go home to her
   parents’ house.’

The viewing arrangement is sketched below in Figure 2. This scene contains two
sequential and identical events TR and LM, where the scope of predication is set on the
entire event chain. Because the speaker is talking about a recurrent event over a period
of time, the gap between the two events is minimized as a trivial one, rendering a
continuous reading.

Figure 2: The viewing arrangement of nahan ‘still’

Example (5) is an excerpt taken from a piece of frog story (Slobin 1996). Note that
nahan is now glossed as ‘first,’ marking the sequentially earlier event of an event series
in its holistic context, an ‘episode.’ In (5), the two sequential activities, ‘the boy took a
look at the container to see his frog,’ and ‘he then found his frog disappeared,’ are
linked by nahan to show that they are temporally relevant to the episode ‘taking a look
in order to find.’ The lexical item prepares the hearer to expect a series of event
temporally ordered and related.

(5) Frog 6
   … korkoring k<om>ita’ nahan ila hini’ ray==
   child <AF>see first PFV this LOC
   … k-’<in>rwasek-an noka== takem
   KA-<PFV>live-NMZ GEN frog
   …(0.9) ro- takem ‘oka’ ila i kita’-i
   FS frog NEG PFV FIL see-PF
   ‘When he woke up, he couldn’t find the frog in the container.’
Example (6), extracted from a piece of pear story (Chafe 1980), displays a similar use of *nahan*. A boy rode away on his bike with a basket of fruits he had stolen, and he suddenly fell on the ground by bumping into a stone. He thus lost his hat on the ground, and rested ‘a while’ by the side of the road to wait for one of those who ran behind to hand him his hat.

(6) Pear5
72. ...(1.0) ta m-intani’ ila *nahan*
    EXHORT AF-stay.on.the.side PFV a.while
    ‘The boy stopped for a while.’
73. ...(1.3) sa-boway-en ila nisia ka== ...’a tatpo’
    go-give-PF PFV 3SG.GEN NOM FIL hat
74. ...(1.1) sa-bow ay-en ka tatpo’
    go-give-PF NOM hat
    ‘(One of the boys) walked toward (the boy) and gave the hat (back to him).’

Example (5) and (6) share the same viewing arrangement. The construal, provided as Figure 3, contains two events sequentially related as a meaningful episode, with the predication scope on both of them. The construal is different from Figure 1 and 2 in that TR is not the repetition of LM. For example, in (6), ‘the boy stopping for a while’ and ‘another boy moving towards’ are not identical as what we have seen in cases of ‘still,’ ‘again,’ and ‘another.’ The TR and LM are represented by two different icons in Figure 3.

![Figure 3: The viewing arrangement of *nahan* ‘first’ and ‘a while’](image)

Figures 1-3 show that the viewing arrangements for the various uses are in fact represented by fairly similar configurations. Those variations are interrelated as a semantic ‘family,’ and should not be regarded as so diverged as what we have assumed by five different English words used to translate them. If Principled Polysemy were adopted, those different renditions of *nahan* are shown to be closely interrelated because they share the same protoscene, as in Figure 4.
This structure expresses an abstract notion of CONTINUATION in which the trajectory is understood against a landmark in a temporal sense. Perspectivization, the angle from which we view a scene (Graumann and Kallmeyer 2002, Langacker 1999), plays a key role here. When we take a narrower perspective to focus on the trajectory only, the interpretation is ‘again’ or ‘another.’ When we take a wider perceptive to view the TR-LM cluster, the trajectory is conceived as an extension of the landmark, hence the reading ‘still.’ For ‘first’ and ‘a while,’ the speaker takes an even wider perspective for the entire episode, denoting event sequence, but not necessarily the repetition effect. In section 4, we will show how different perspectives are constrained and motivated, and how the various meanings are in fact chained to each other as a family, if viewed as different perspectivizations of the same construal.

4. Semantic extension in discourse

To explain the mismatch between semantic meaning and pragmatic interpretation, Langacker (1999) suggests that meaning interpretation is not autonomous. Instead, it is grounded in its speech interaction, and he thus proposes ‘grounding,’ which includes the speech event, its participants, and the surrounding context, to handle this mismatch, as well as Current Discourse Space (CDS) to manipulate meaning inference (Langacker 2002).

What Langacker has advanced is a big step toward the form and meaning pairing, which can reflect the very nature of language if he could make explicit what he calls ‘context.’ In this section, we would like to argue that linguistic collocation, the discourse, and the speech situation play significant roles in meaning inference. Such inference is indispensable in the understanding of the polysemous nahan in Saisiyat, and also the issue of meta-linguistic glossing in general.
4.1 Textual effect

Two types of textual effects are at issue here. The first type concerns the coercion effect of the collocating words, and the other one concerns how the preceding discourse may affect meaning interpretation.

Collocation effect can be illustrated by the link between ‘again’ and ‘another.’ The difference between ‘again’ and ‘another’ lies in that one is used in temporal context, and the other in the listing context. In the listing context, nahan is collocated with numeral predicates, and it is this nominal predicate that provides the context to yield the coerced meaning of ‘another’ in English.

The collocating effect is also found in the link between ‘again’ and ‘still.’ Vendler (1967) proposed that verbs can be categorized into four types with respect to their ‘event structure.’1 Among them, ‘accomplishments’ and ‘achievement’ hint at event termination whereas ‘states’ and ‘activities’ do not.2 We found that the meaning ‘again’ apply to ‘accomplishments,’ ‘achievements’ and ‘activities,’ whereas the meaning ‘still’ is prone to be activated by ‘states’ and some ‘activities.’ For example, in (7), an accomplishment event ‘to climbing to the top of the tree’ implies a termination where the ‘continuous’ reading should be blocked. The ‘still’ reading is thus excluded.

(7) Pear 3
8. ...(1.7) insa’an kasna’itol ray kahoey babaw nahan
now AF.move.up LOC tree above again
‘(The old man) went up the tree again.’

Similarly, (8) depicts a state in which the old man is picking up fruits on the tree, and the event type of the state invites a ‘still’ reading.

---

1 Also known as ‘eventuality types,’ ‘aspects,’ ‘aspectual classes,’ ‘inherent lexical aspect,’ or ‘Aktionsart’ (Filip 1999:15).
2 Many modern linguists come to realize that Achievements and Accomplishments are overlapped in terms of linguistic phenomena, and States and Activities (or Processes) in some respects exhibit significant semantic and syntactic similarities. Two super-groups are hence distinguished: the Event group and the State/Process group. The former group includes telic and quantized verbs while the latter includes atelic and cumulative verbal predicates (Filip 1999:16).
(8) Pear 3

44. ...(1.2) isaza tatini’ rima’ r<om>okrok
   FIL old.man AF.go <AF>pick
   nahan babaw ka boway
   still above ACC fruit

   ‘The old man was still up in the tree to pick fruits.’

However, this classification scheme can be ‘fuzzy’: the distinction between event types is not necessarily responsible for the different readings. Linguists begin to recognize recently that events are distinguished according to our experience rather than determined solely by language, and verbal categories are not separated by clear-cut boundaries (Filip 1999). A verbal predicate might have in general a primary category, but it might be classified, via contextual coercion, as another verbal category. State predicates like ‘al’alak ‘young’ in (2) can yield the ‘again’ reading when the situation so requires. In general, collocations play a less significant role in meaning interpretation.

The second type of textual effect on meaning refers to the influence of the preceding linguistic context, which we will illustrate with (9). In IU30-32, the speaker said a group of people came back after hunting, which would affect the interpretation of IU62, when the speaker talked about their hunting the next day. Here, nahan refers to the hunting act is repeated and is interpreted naturally as ‘again.’

(9) Kathethel 2

30. ...(1.2) lasia ma’ isaa==
   3PL.NOM also that
31. ... rima’ ila ‘<oem>alep
   AF.go PFV <AF>hunt
32. ... lobih
   AF.return

   ‘They went back from hunting.’

…

60. ...(1.5) m<-in>‘itol ‘aehae’ hahila
   AF<-PFV>stand one day
61. ... lasia
   3PL.NOM
62. ...(0.8) rima’ nahan ‘<oem>alep
   AF.go again <AF>hunt

   ‘On the next day, they went hunting again.’
4.2 Situational effect

Situational effect refers to influence of the speech circumstances on our understanding and meaning interpretation. In §4.1, we have illustrated that *nahan* is interpreted as ‘again’ due to textual constraint. In this section, we shall see how situation may play a role in inducing the ‘still’ reading of *nahan*, as shown in (10). Speaker M in (10) asks whether Hiziyo’ hunts when the two speakers talk about their daily chores. Knowing that Hiziyo’ has hunted habitually for a period of time and such habitual action has not been interrupted, M’s use of *nahan* generates the ‘still’ reading in accordance to the background information of the situation.

(10) Life
151. M: ...(1.4) rangi’ hiziyo’ rima’ *nahan* ray ... bangol ... talboeyoe’
   VOC PN AF.go **again** LOC forest AF.hunt
   ‘Does Hiziyo’ **still** go hunting?’
152. F: ...(1.7) ‘oka’ ila
   NEG PFV
   ‘No.’
153. M: ... ‘oka’ ila
   NEG PFV
   ‘No.’
154. F: m==
   FIL
155. F: ...(0.9) sia mamowa’ ila ka ka’niw
   3SG.NOM AF.plant PFV ACC mushroom
   ‘He plants mushroom.’

Situational effect has in fact been pointed out by studies of categorization. Ungerer and Schmidt (1996) maintain that prototype is situated in the context, as illustrated by (11):

(11) He opened the door to face a pretty young woman with a *dog* in her arms.

The kind of dog that first comes to our mind in this situation is probably a small lapdog, e.g., Pekingese, though a golden retriever may be more prototypical for the concept of ‘dog.’ When the situational context is taken into consideration, our understanding of a lexical category will be modified so that our interpretation will make sense, so as to reduce potential ambiguity or vagueness in any natural, cooperative communication.
The ‘first’ and ‘a while’ readings in (12) are largely motivated by situational context. *Nahan* is used for the two activities mentioned in (12): ‘eating at Ayon’s home’ in IU 231 and IU232 and ‘going to watch workers planting ginger’ in IU233. The interpretations ‘first’ and ‘again’ are assigned respectively to describe the two activities as a meaningful episode.

(12) Life
231. F: ... (1.1) hay komsi’ael **nahan** ray taewan
   LOC have.lunch **first** LOC house
   ‘You ate lunch at Ayon’s home first.’
232. M: ... ‘ihi’ komsi’ael **nahan** taewan o:
   BC have.lunch **first** house DM
   ‘Yes. I had lunch at her house first.’
233. M: .. rima’ **nahan** k<om>ita’ lasia ‘am==
   AF go again <AF> see 3PL.NOM FUT
234. M: ...(0.8) somobaboeyoe ha:w
   prepare.soil.for.planting.ginger there
   ‘(After lunch), I went to see the workers preparing to plant ginger there.’

The meaning ‘first’ is an elaboration of the situation. The protoscene predicts a chain of sequentially related events. Since IU 231 is the first element of the event sequence, the informant imposes the ‘first’ or ‘a while’ reading to make sense of the situation expressed by the utterance.

4.3 An interrelated semantic network

The figure below shows the semantic network of Saisisyat *nahan*. In fact, those meanings are so closely related that similarities between any two single nodes in the chart can be easily identified. The tight chain connecting the various uses makes it hard for us to find the kind of discreteness common in a case of typical polysemy. Modulated or coerced meanings may come, as discussed above, from situational and textual contexts.
With the protoscene in mind, we therefore propose that contextual variations should not be treated as separate senses that are stored in our mental lexicon. The implication of this proposal is that there should be some way to distinguish core meaning from its pragmatic extensions, when it comes to meta-linguistic glossing.

5. Cultural context

In meta-linguistic glossing, we often encounter difficulty in assigning in the target language appropriate ‘equivalent’ for a lexical item. Such cross-linguistic difficulty often involves, among other things, cultural factors. The following example from Tsou, also a Formosan language, clearly shows the problem of this nature.

\[(13)\] mako ahtu uh ne hnou ho bonu no f’ue
\hspace{1cm} go Obl hut Conj eat Obl yam
\hspace{1cm} ‘You can come to my house and enjoy a meal.’

Two interesting questions arise from (13): (a) why do the speakers use the word \textit{hnou} ‘hut’ to refer to one’s house? And, (b) why do they use the plant \textit{f’ue} ‘yam’ as something to offer and also to hint at the shabbiness of their offer?

Language reflects the living environment and the social construct of its users, so much so that the metaphorical reading of ‘a humble dinner’ is expressed through whatever is available and probably ordinary to the people of the language. It should follow then that the Tsou people primarily live, at least at one point in history, in huts
and consume the yam as a common-place food.³

What we are suggesting here is that the extensively discussed notion of ‘context’ should include cultural knowledge. The notion of context has grown in a rapid manner (Ungerer and Schmidt 1996) mostly in view of the insufficiency of analyzing language as decontextualized patterns by generative linguistics. The traditional definition of taking context as the ‘linguistic material’ preceding or following a word or a sentence is revolutionalized when Searle (1979:25) expends it as any background assumptions ‘that are necessary for an utterance to be intelligible.’ In recent discourse studies, contexts are defined as any information related to the situation in which an utterance is embedded, which is the broad definition we would like to embrace here.

The prototypicality of a category is also discussed in literature on categorization as greatly influenced by cultural context. For example, Lipka’s (1987) experiment shows that the prototypical DESK in Japan and China is with shorter legs, or without drawers, because of the traditional shape associated in the two cultures. It is therefore suggested that the internal structure of a category could differ from culture to culture. To understand a lexical category desk and the possible meaning of it in the language it is used, we must have relevant knowledge of that culture. The gender assignment of nominals in Northern Iroquoian language can be used as another example (Chafe 2002). According to Chafe’s study, speakers of that language mark the third person pronoun as masculine, feminine and neuter based on their cultural model of the gender stereotypes. In Northern Iroquoian culture, masculine is depicted as conspicuous, flamboyant and independent, whereas females as stable, undifferentiated and inconspicuous.

Even though Malinowski has suggested the notion such as ‘context of culture’ to advocate the cultural aspects of language, few studies are devoted to this endeavor (Halliday and Hasan 1989). Diversity can be expected if different linguistic communities draw on different experiential bases for their conceptualization of reality. Reference to this experiential base of a lexicon often stimulates meaningful discussion of a question that is often raised in conjunction with the relationship between language and culture. Whether the dichotomy is arbitrarily determined or a motivated one remains an issue to be explored in greater detail.

6. Implications to semantic fieldwork

The research on the use of meta-language for semantic glossing (Matthewson 2004, Lehrer 1992) should therefore be given high priority, though there are challenges

³ According to Yeh (2001), Tsou traditional buildings are constructed by woods and bamboo. The roof is covered by thatch. The traditional everyday foods of Tsou are millet and yam.
against this ‘necessary evil,’ claiming that the semantics of a lexicon can be distorted when represented by the second language. It is true that Chindali has nine parameters in constructing their verbs of motion which cannot be simply glossed as ‘come’ and ‘go’ (Botne 2005), we nevertheless find the texts hardly conceivable if those verbs are glossed by schematic symbols unnatural or even counter-intuitive to speakers of the language. We maintain that corpus should be designed in such a way that is readable and accessible to its users. The problem is how we can adequately represent the specific semantic categorization of an object language with meta-linguistic glossing.

A typical text transcription follows the three-line format, as illustrated below:

(14) Line 1: Object language  
     Line 2: Morph-by-morph gloss  
     Line 3: Free meaning translation

We have already spelled out the two considerations which make the direct translation approach a questionable practice, the two problems being the language-specific categorization of a conceptual space, and the distinction between semantic core and its contextual extensions. Semantic glossing should be precise enough to capture the semantic categorization of a language, yet flexible enough to reflect the rich extensions in pragmatic concern.

We thus advance Hypothesis I below concerning the necessary distinction between sense and force:

_Hypothesis I: Line 2 should be saved for semantic core that is the prototypical meaning of that lexical item. Pragmatic extensions should be indicated in Line 3, which is reserved for free translation situated in context._

In the case of Saisiyat _nahan_, a unifying gloss like CONTINUE could be used in Line 2 over the five different English translated equivalents. This approach helps corpus users to understand that these are related uses of the same cognitive category. Semantic overlapping (family resemblance) and contextual elaborations are to be expected. Pragmatic elaborations are to be documented in Line 3, to make sense of the linguistic form in context.
432

Lily I-wen Su and Shuping Huang

(4) repeated

76. ...(1.1) hiza minkoringan ‘am==
     that woman FUT

77. .. lobih k<om>osa ‘am rima’ kala yaba nahan
     AF.return <AF>say FUT AF.go LOC father CONTINUE

78. ... nisia
     3SG.GEN

79. ... rima’ ray ‘inkahoelan
     AF.go LOC married.woman’s.parents’.home

‘(After getting married for a long time), the woman still wanted to go home to her parents’ house.’

The tree-line convention of text glossing also represents two different views on the nature of cross-linguistic communication, as expressed in our Hypothesis II, which concerns the relativist or the universalist views inherent in corpus transcription:

Hypothesis II: Those who believe Line 2 as the proper translation stand for a cultural relativist approach, whereas those who believe Line 3 as the proper translation stand for the universalist approach.

Line 2, the morph-by-morph gloss, is the place that directly reflects the mapping of form and meaning in a language. In the belief of linguistic relativism, meaning of a language lies in its formal structure. When forms diverge from one language to another, meanings are supposed to be culturally relative (Boas 1940, Whorf, cited from Carroll 1956). The semantic counterparts given by the informants are too often found to be irrelevant to the core meaning of a lexical item in the object language, primarily because it reflects social-cultural background of the speech community hardly translatable. In his investigation of lexicons in Wayan Fijian dialect, Pawley (2006) finds the dialect rich in terms for sea animals organized in systematic taxonomic structure and polysemous extension. This is so because Wayan Fijians earn their living largely by catching edible sea animals. A Wayan Fijian lexicon, say, ika ‘fish’ can in fact mean ‘fish,’ ‘whales,’ ‘dolphins’ and even ‘turtles.’ To gloss ika in line 2 as ‘fish’ and let the instantiations be represented in line 3, can profile its cultural-specific way of conceptualizing certain sea animals.4

For the universalists, those advocates of universalism, it is suggested that line 3 be taken as the proper translation of the object language. They maintain that meaning

4 Here we do not propose a strong version of linguistic determinism, but simply point out cross-cultural difference of language use.
cannot be fully embedded in lexicon or grammar, because forms of a language are constrained by the linguistic repertoire highly conventionalized or formalized. Unlike the relativists who tend to over-estimate the power of social environments, the universalists neglect the structural difference, assuming our ability to understand meanings in its context, and represents more or less a belief in the universality of language use (cf. Wierzbicka 1996, Goddard 2001). The universalists believe that it is the translation in line 3, rather than grammar, that affects our cognition.

We therefore hold that cultural-specific knowledge should be preserved in corpus documentation. This knowledge contains precious information that is of particular value to the study of endangered languages. The problem is how such knowledge can be effectively represented and documented. It is our hope that we can include information under the so-called ethnolinguistic notes in our corpus. We will use the following example to illustrate what we have in mind.

(15) a. Life

9. F: ...(2.4) t<om>angtang ay
<AF>dig Q
‘Do you do any digging?’
10. M: .. eh==
DM
11. F: ((CLEARS THROAT))
12. M: ...(0.9) t<om>angtang noka== .. kama-si’ael ka samiyan
<AF>dig use KAMA-AF.eat ACC god
‘I use a machine.’

b. Flood

28. ...(7.4) kayzaeh ila p<in>askayzae==h
AF.good PFV <PFV>AF.do
‘Finally, (the boat) was completed.’
29. ... samiyan m-wa:i’ ila k<om>ita==‘
god AF-come PFV <AF>see
‘And the god came to see it.’

c. Molaw

36. ... nak’isaa
like.that
37. .. hayza’ ‘aehae’ hahila
exist one day
38. ..(0.8) ‘inaray babih
from another.place
Speaker F in (15a) asks whether M digs holes to plant gingers. If we maintain the semantic information in Line 2 and pragmatic extension in Line 3, what IU 12 literally means ‘(I) use the thing that eats god.’ will be all confusing, if no possible link is provided to make sense of the utterance. The term samiyan, originally denoting ‘god’ in Saisiyat, as in (15b) and (15c), is used here to refer to any entity with super-natural or unexplainable power, for instance, gasoline. Gasoline can be something unexplainable and even super-natural for the non-industrialized Saisiyat society. A note explaining this special link would be helpful in providing the necessary cultural background, and to bridge the gap between Line 2 and Line3.

In summary, precision is not the only key to the methodology of corpus documentation, appreciation of its culture should also be taken into consideration. In developing a corpus for Austronesian languages, especially the endangered Formosan ones, our methodological concern is summarized in the Table below:

<table>
<thead>
<tr>
<th>Content</th>
<th>Component</th>
<th>Ideology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 1 Object language</td>
<td>form</td>
<td>culture relativity</td>
</tr>
<tr>
<td>Line 2 Morph-by-morph gloss*</td>
<td>sense</td>
<td>function universality</td>
</tr>
<tr>
<td>Line 3 Free translation</td>
<td>force</td>
<td></td>
</tr>
</tbody>
</table>

*Ethno-linguistic notes context

7. Conclusion

In developing a corpus of Austronesian languages, we have encountered several empirical problems that prompt our investigation into meta-linguistic glossing and semantic theories. Via ideas from the framework of Principled Polysemy, we suggest that the problems be handled from the perspective of linguistic categorization.

Meaning, from a language internal viewpoint, is often situated in context, which has significant bearing on the semantic divergence away from its protoscene. Manipulated by cultural-specific knowledge, interpretation of the lexical item and its semantic extensions...
are largely dependent upon the conceptualization of a cognitive model in that specific culture, a point well made by Taylor (1995:132), ‘All meaning is, in a sense, pragmatic, as it involves the conceptualizations of human beings in a physical and social environment.’ The study of semantics is hence closely-intertwined with pragmatic considerations, and it is not economical to enlist every use of a lexical item whenever it is used in a novel context (Riemer 2001).

From a cross-linguistic point of view, glossing the object language with a meta-language often entails overlooking the mismatch of semantic categorization between the two languages, as shown by *nahan* in Saisiyat. Mithun (2001) warns that research methodologies can shape the resulting data in ways that the researchers are not even aware of. Meta-linguistic glossing often transfers in an implicit manner the lexical categorization of the meta-language into the object language.

Meta-linguistic glossing, an inevitable component in corpus building, should be designed in such a way that reflects the methodological considerations mentioned above. It is hoped that notions that have been widely advocated in categorization studies can shed lights on corpus documentation.
References


Slobin, Dan I. 1996. Two ways to travel: verbs of motion in English and Spanish.
葉春榮 (Yeh, Chuen-rong) 主編. 2001. 《番族慣習調查報告書第四卷：鄒族》。台北市：中央研究院民族學研究所。（編譯自 1918 年台灣總督府「臨時台灣舊慣調查會」以日文出版之《番族慣習調查報告書第四卷：鄒族》）