Quantifying semantic shift for reconstructing language families

William Croft Clayton Beckner Logan Sutton
University of New Mexico

Jon Wilkins Tanmoy Bhattacharya Daniel Hruschka
Santa Fe Institute

Abstract

In comparative historical linguistics, one must weigh evidence from large numbers of putative cognates in order to arrive at the best hypothesis of the family tree and reconstructions. The comparativist presently uses unquantified knowledge of these processes. We present a typological study of word polysemy in order to construct a quantified network of semantic similarity among basic vocabulary items for comparative historical research. We investigate 22 concepts denoting natural objects in the Swadesh list across a typological sample of over 50 languages. In addition to its value for comparative historical linguistics, the study also reveals universals of lexical conceptual space.

1. Introduction

The need

'…historical linguistics cannot ignore semantic change. For unless we can relate words such as Old English *hlāf* 'bread' and New English *loaf* not only phonetically but also semantically, it is impossible to trace many historical developments and to do meaningful historical linguistic research' (Hock 1986:284).

The problem

'there seem to be no natural constraints on the directions and results of semantic change. Given enough imagination—and daring—it is possible to claim semantic relationship for almost any two words under the sun.' (Hock 1986:308)

'There is…little in semantic change which bears any relationship to regularity in phonological change' (Fox 1995:111)

The status quo

'If the correspondences are regular, the set of words is cognate, however unlikely the semantics. That is, structural grounds—regular correspondences—are sufficient for establishing cognacy, while semantic grounds are neither necessary nor sufficient. (Nichols 1996:57, describing a 'working assumption' of the comparative method)

A better way

*What matters is not **possibility**—yes, anything can happen—but **probability**—the greater or lesser likelihood of semantic shift to take place. This requires **quantification** on an **empirical** basis.*

2. An example: Eskimo-Aleut and Dravidian 'night'

•In Eskimo-Aleut (Fortescue, Jacobson and Kaplan 1994:373), Yupik 'night' is equated with Inuit 'evening'; this equation is necessary to link these two branches
•Eskimo 'night'/'evening' is equated with an Aleut word 'long ago' which appears to be a negative suffix attached to a stem glossed as possibly 'tonight'
•In Dravidian (Burrow and Emeneau 1984, etymology 2552), South Dravidian forms meaning 'night' and 'darkness' are equated
•South Dravidian 'night'/'darkness' is equated with Central Dravidian forms meaning 'charcoal, coal, soot'

3. How can we quantify semantic relationships?

Direct observation

•Survey documented semantic changes (Williams 1976, Sweetser 1990, Traugott & Dasher 2003, and many grammaticalization studies—e.g., Heine & Reh 1980, Lehmann 1982/1997, Heine & Kuteva 2002).
•*Problem:* there are very few languages documented across time, and they do not form a good sample of the world's languages. Also, present research has focused on grammaticalization, not lexical semantic change

Reconstructed changes

•Survey reconstructed semantic changes in various language families, as has been done for body-part terms (Wilkins 1996)
•*Problem:* presupposes what we are trying to discover—it is based on linguist's intuitions of what are plausible semantic relationships; cannot be quantified

Typology of polysemy

•Survey polysemy patterns in a synchronic typological sample.
•The first step in a semantic change is extension of a word to a new meaning
•A crosslingistic sample will allow us to quantify the likelihood of semantic change in a particular time slice (represented by the synchronic description)

Antecedents

•Typological analyses of polysemy have been made of domains such as perception verbs (Viberg 1983) and color metaphors (Derrig 1978); they do not discuss the implications for comparative historical linguistics.
•Evans (1992) surveys polysemy of concepts associated with FIRE, and Evans & Wilkins (2000) surveys the polysemy of perception/cognition verbs, in Australian Aboriginal languages, with the goal of aiding comparative historical linguistics; they focus on culture-specific polysemies (see our Results and Discussion).
•Brown and colleagues do typological studies of the polysemy of body-part terms, which form an important part of the basic vocabulary used for comparative historical linguistics (Swadesh 1952, 1955; see Brown 1976, 1979; Brown et al. 1976; Witkowski & Brown 1978; Brown & Witkowski 1981) and also cardinal direction terms (Brown 1983) and deictic terms (Brown 1985). These studies are the most direct antecedents to ours. Brown et al. do not focus on the ability to quantify semantic shift.
•Brown et al., Evans & Wilkins and Viberg start from a set of concepts and survey their expression across languages. Derrig and Evans start from a set of concepts and surveys what other concepts are expressed with the same word forms. Ideally, both methods should be used in combination. In the pilot study reported here, we follow Derrig and Evans; in future work, we will survey the expressions of all the concepts with the strongest links to the initial concept set.

4. Methods

Selecting the sample

•81 languages from different genera (Dryer 1989) with a broad geographical distribution, using available, good quality dictionaries (see Appendix).
•22 concepts from the Swadesh basic vocabulary lists referring to physical entities other than body parts:

Celestial Phenomena	**Natural Substances**	**Landscape Features**
SUN	SALT	SEA/OCEAN
MOON	WIND	LAKE
STAR	SMOKE	RIVER
NIGHT	WATER	MOUNTAIN
DAY/DAYTIME	FIRE	SKY
YEAR	ASH(ES)	CLOUD(S)
	STONE/ROCK	
	SAND	
	EARTH/SOIL	
	DUST	

Finding polysemies

•All word forms expressing these concepts were identified in the dictionaries
•Polysemies for the word forms were then identified.
•Meanings under separate entries were generally excluded.

•Borrowed words were included, as well as derived or otherwise analyzable expressions.
•An English word used in the analysis may correspond to a set of near-synonyms in the actual dictionaries. For example, we grouped together semantically similar translation equivalents (e.g. 'sunlight', 'sunshine', 'daylight'; 'pile, heap, mound'; 'stream, brook, creek').

5. Results and discussion

Historical linguistics

•There are clearly major differences in the probability of semantic shift for different concepts which are not intuitively obvious. Hence typological studies of polysemy can provide empirically justified quantification of degrees of semantic similarity, which allows one to evaluate probabilistically hypotheses of cognacy among words that are not translation equivalents (see the Postscript below right)

Semantics

•The study also has consequences for understanding the nature of semantic change. For example, what appears to be the "same" metonymic shift does not have the same frequency in different cases:

– TIME INTERVAL ⇔ BEGINNING OF INTERVAL "explains" NIGHT ⇔ EVENING and DAY/DAYTIME ⇔ DAWN. But the former is far more frequent than the latter.
– CELESTIAL OBJECT ⇔ TIME PERIOD OF CYCLE "explains" MOON ⇔ MONTH and SUN ⇔ DAY (or SUN ⇔ YEAR). But the former is far more frequent than the latter.
– CELESTIAL OBJECT ⇔ LIGHT EMITTED BY OBJECT "explains" SUN ⇔ SUNLIGHT/DAYLIGHT, MOON ⇔ MOONLIGHT and STAR ⇔ STARLIGHT. But the last one is much less frequent than the other two.
– On the other hand, LARGE LANDSCAPE FEATURE ⇔ SMALLER LANDSCAPE FEATURE, a type of semantic extension or generalization, does appear to occur with similar frequency among MOUNTAIN ⇔ HILL, LAKE ⇔ POND, and RIVER ⇔ STREAM/BROOK/CREEK.

Typology and universals

•Cultural and ecological factors appear to play a role in influencing certain polysemy patterns (cf. Witkowski, Brown & Chase 1981; Brown & Witkowski 1983; Witkowski & Brown 1985; Evans & Wilkins 2000)

–SUN ⇔ MOON occurs only in North America
–STAR ⇔ LUCK is found only in the Middle East area

•Hence language universals will need to incorporate cultural and ecological properties in their formulation

6. P.S. So what about Eskimo-Aleut and Dravidian 'night'?

•Eskimo NIGHT ⇔ EVENING is a highly likely equation, so the semantics supports these cognates
•Aleut suggested etymology (not) TONIGHT ⇔ NIGHT is not supported by the

typological survey. A more plausible candidate would be LAST NIGHT or 24HR PERIOD ⇔ NIGHT.

•Dravidian NIGHT ⇔ DARKNESS is a quite likely equation, so the semantics supports these cognates

•Dravidian NIGHT ⇔ CHARCOAL/COAL/SOOT is not supported by the typological survey; in fact, CHARCOAL/COAL/SOOT (symbolized by EMBERS) is likely to be equated with FIRE or SMOKE

7. Appendix: languages in the sample

Region	Family*	Genus	Language
Africa	Khoisan	Northern	Ju|'hoan
		Central	Khoekhoegowab
		Southern	!Xoó˜
	Niger-Kordofanian	NW Mande	Bambara
		Southern W. Atlantic	Kisi
		Defoid	Yoruba
		Igboid	Igbo
		Cross River	Efik
		Bantoid	Swahili
	Nilo-Saharan	Saharan	Kanuri
		Kuliak	Ik
		Nilotic	Nandi
		Bongo-Bagirmi-Kresh	Kaba Deme
	Afroasiatic	Berber	Tumzabt
		West Chadic	Hausa
		E Cushitic	Rendille
		Semitic	Iraqi Arabic
Eurasia	Basque	Basque	Basque
	Indo-European	Armenian	Armenian
		Indic	Hindi
		Albanian	Albanian
		Italic	Spanish
		Slavic	Russian
	Uralic	Finnic	Finnish
	Altaic	Turkic	Turkish
		Mongolian	Khalkha Mongolian
	Japanese	Japanese	Japanese
	Chukotkan	Kamchatkan	Itelmen (Kamchadal)
	Caucasian	NW Caucasian	Kabardian
		Nax	Chechen
	Kartvelian	Kartvelian	Georgian
	Dravidian	Dravidian Proper	Badaga
	Sino-Tibetan	Chinese	Mandarin
		Karen	Karen (Bwe)
		Kuki-Chin-Naga	Mikir
		Burmese-Lolo	Hani
		Naxi	Naxi

Oceania†	Hmong-Mien	Hmong-Mien	Hmong Njua
	Austroasiatic	Munda	Sora
		Palaung-Khmuic	Minor Mlabri
		Aslian	Semai (Sengoi)
	Daic	Kam-Tai	Thai
	Austronesian	Oceanic	Trukese
	Indo-Pacific/Papuan	Middle Sepik	Kwoma
		E NG Highlands	Yagaria
		Angan	Baruya
		C and SE New Guinea	Koiari
		West Bougainville	Rotokas
		East Bougainville	Buin
	Australian	Gunwinyguan	Nunggubuyu
		Maran	Mara
		Pama-Nyungan	E and C Arrernte
Americas	Eskimo-Aleut	Aleut	Aleut
	Na-Dene	Haida	Haida
		Athapaskan	Koyukon
	Algic	Algonquian	Western Abenaki
	Salishan	Interior Salish	Thompson Salish
	Wakashan	Wakashan	Nootka (Nuuchahnulth)
	Siouan	Siouan	Lakhota
	Caddoan	Caddoan	Pawnee
	Iroqoian	Iroquoian	Onondaga
	Coastal Penutian	Tsimshianic	Coast Tsimshian
		Klamath	Klamath
		Wintuan	Wintu
		Miwok	Northern Sierra Miwok
	Gulf	Muskogean	Creek
	Mayan	Mayan	Itzaj Maya
	Hokan	Yanan	Yana
		Yuman	Cocopa
	Uto-Aztecan	Numic	Tümpisa Shoshone
		Hopi	Hopi
	Otomanguean	Zapotecan	Quiavini Zapotec
	Paezan	Warao	Warao
		Chimúan	Mochica/Chimu
	Quechuan	Quechua	Huallaga Quechua
	Araucanian	Araucanian	Mapudungun
	Tupi-Guaraní	Tupi-Guaraní	Guaraní
	Macro-Arawakan	Harakmbet	Amarakaeri
		Maipuran	Piro
	Macro-Carib	Carib	Carib
		Peba-Yaguan	Yagua

*Families vary as to degree of acceptance.
†Includes families in Southeast Asia.

8. References

Brown, Cecil H. 1976. General principles of human anatomical partonomy and speculations on the growth of partonomic nomenclature. *American Ethnologist* 3.400-24.

Brown, Cecil H. 1979. A theory of lexical change (with examples from folk biology, human anatomical partonomy and other domains). *Anthropological Linguistics* 21.257-76.

Brown, Cecil H. 1983. Where do cardinal direction terms come from? *Anthropological Linguistics* 25.121-61.

Brown, Cecil H. 1985. Polysemy, overt marking, and function words. *Language Sciences* 7.283-322.

Brown, Cecil H., John Kolar, Barbara J. Torrey, Tipawan Trương-Quang and Phillip Volkman. 1976. Some general principles of biological and non-biological folk classification. *American Ethnologist* 3.73-85.

Brown, Cecil H. and Stanley R. Witkowski. 1981. Figurative language in a universalist perspective. *American Ethnologist* 8.596-615.

Brown, Cecil H. and Stanley R. Witkowski. 1983. Polysemy, lexical change and cultural importance. *Man* (N.S.) 18.72-89.

Burrow, Thomas and Murray B. Emeneau. 1984. *Dravidian etymological dictionary*, 2nd ed. Oxford: Oxford University Press.

Derrig, Sandra. 1978. Metaphor in the color lexicon. *Papers from the Parasession on the Lexicon, Chicago Linguistic Society*, ed. Donka Farkas, Wesley M. Jacobsen, and Karol W. Todrys, 85-96. Chicago: Chicago Linguistic Society.

Evans, Nicholas. 1992. Multiple semiotic systems, hyperpolysemy, and the reconstruction of semantic change in Australian languages. *Diachrony within synchrony: language history and cognition*, ed. Günter Kellermann and Michael D. Morrissey, 475-508. Frankfurt: Peter Lang.

Evans, Nicholas and David Wilkins. 2000. In the mind's ear: the semantic extensions of perception verbs in Australian languages. *Language* 76.546--92.

Fortescue, Michael, Steven Jacobson & Lawrence Kaplan. *Comparative Eskimo dictionary, with Aleut cognates*. Fairbanks: Alaska Native Language Center.

Fox, Anthony. 1995. Linguistic reconstruction: an introduction to theory and method. Oxford: Oxford University Press.

Heine, Bernd & Tania Kuteva. 2002. *World lexicon of grammaticalization*. Cambridge: Cambridge University Press.

Heine, Bernd & Mechthild Reh. 1984. *Grammaticalization and reanalysis in African Languages*. Hamburg: Helmut Buske Verlag.

Hock, Hans Heinrich. 1986. *Principles of historical linguistics*. Berlin: Mouton de Gruyter.

Lehmann, Christian. 1982/1995/2002. *Thoughts on grammaticalization: a programmatic sketch, Vol. I*, Arbeiten des Kölner Universalien-Projekts, 48, Köln: Institut für Sprachwissenschaft. Revised edition published by LINCOM Europa, München, 1995. Revised edition reprinted as Arbeitspapiere des Seminars für Sprachwissenschaft der Universität Erfurt, 9. Erfurt: Seminar für Sprachwissenschaft der Universität.

Nichols, Johanna. 1996. The comparative method as heuristic. *The comparative method reviewed: regularity and irregularity in language change*, ed. Mark Durie and Malcolm Ross, 39-89. Oxford: Oxford University Press.

Swadesh, Morris. 1952. Lexico-statistical dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96.452-63.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistical dating. *International Journal of American Linguistics* 21.121-37.

Sweetser, Eve. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

Traugott, Elizabeth and Richard Dasher. 2002. *Regularity and semantic change.* Cambridge: Cambridge University Press.

Viberg, Åke. 1983. The verbs of perception: a typological study. *Linguistics* 21.123-162.

Wilkins, David. 1996. Natural tendencies of semantic change and the search for cognates. *The comparative method reviewed*, ed. Mark Durie and Malcolm Ross, 264-304. Oxford: Oxford University Press.

Williams, Joseph M. 1976. Synaesthetic adjectives: a possible law of semantic change. *Language* 52.461-478.

Witkowski, Stanley R. and Cecil H. Brown. 1978. Lexical universals. *Annual Review of Anthropology* 7.427-51.

Witkowski, Stanley R. and Cecil H. Brown. 1985. Climate, clothing and body-part nomenclature. *Ethnology 24*.197-214.

Witkowski, Stanley R., Cecil H. Brown and Paul K. Chase. 1981. Where do tree terms come from? *Man* (N.S.) 16.1-14.