Exemplar semantics

William Croft, *University of New Mexico (wcroft@unm.edu)*

1. Exemplar theory in phonology

In this paper, I outline what an exemplar-based model of semantics, or more generally, of grammar (the form-meaning pairing) might look like. The suggestions in this paper are therefore quite tentative. I begin by outlining the exemplar approach to phonology, then consider the usage-based model in grammar, and then turn to how the usage-based approach can be made into a more exemplar-based model.

The exemplar approach to phonology has been advocated by Pierrehumbert (2001, 2003) and Bybee (2001). The primary empirical motivation for an exemplar approach to phonology is the well-documented extreme variability in the phonetic realization of phonological categories such as phonemes:

> The phonetic inventory of a language is a set of labeled probability distributions over the phonetic space…The claim that languages use regions of the phonetic space—as opposed to points in the space—is supported by the fact that the phonetic realization of any given element is *always* variable. Even repeated recordings of the same speaker saying the same word in the same context will yield some variability in the measured values of physical parameters (Pierrehumbert 2003:182, 184)

The traditional phonological model posits a single ideal phonetic value for a phoneme, represented by a point in phonetic space. The traditional model therefore ignores the variability

in actual language use, or abstracts away from it as performance errors. The exemplar model on the other hand takes the variability as part of a speaker's knowledge about their language:

> In an exemplar model, each category is represented in memory by a large cloud of remembered tokens of that category. These memories are organized in a cognitive map, so that memories of highly similar instances are close to each other and memories of dissimilar instances are far apart…The entire system is then a mapping between points in a phonetic parameter space and the labels of the categorization system (Pierrehumbert 2001:140)

Thus, the exemplar model defines phonological categories as regions in the phonetic space, defined by a probability distribution that is in turn a generalization over 'a large cloud of remembered tokens.'

Pierrehumbert offers several advantages that the exemplar model has over the traditional single value model (Pierrehumbert 2001:143-44):. The exemplar model accounts for evidence that phonetic lexical detail is remembered and stored by speakers. The exemplar model allows for the modeling of frequency effects, for which there is also strong empirical evidence. A prototype for a phonological category can be defined in terms of the structure of the cloud or cluster of exemplars. Finally, the goodness or extreme examples of the category can be modeled as distance from the modal values of contrasting categories on the phonetic space.

The exemplar model, in this simple form, immediately poses a problem: do we really remember every single token we have ever heard or produced? This seems to be psychologically implausible. But the exemplar model does not really make this assumption. First, memories of individual tokens decay (Pierrehumbert 2001:140). Second, it is hypothesized that the phonetic

space is granular. In other words, some phonetic differences are so fine that they are not distinguished by speakers even in the exemplar model, and so the different tokens are treated as identical (Pierrehumbert 2001:140-41; Bybee 2001:52). Finally, speakers reorganize the representation of words (the locus of phonemic representations). The reorganization of word representations is in part due to frequency effects (Bybee 2001:138-43). But word representations can also be reorganized due to nonphonological effects that have been described by sociolinguists, such as the reallocation of variants and the acquisition of a social value by a variant (see Croft 2000:174-78, and references cited therein).

The exemplar model is still largely programmatic in the area of phonology, although it is a response to well-known empirical facts and involves more than simply storing tokens of phonological categories. Applying the exemplar model to grammar, including semantics, poses many greater problems, not just because the phonological model is still young but also because of additional complexities in the pairing of form and meaning. Nevertheless, some important elements of the exemplar model can be found in the usage-based model which has been advocated by cognitive linguists such as Bybee and Langacker for many years (Bybee 1985, 2001; Langacker 1988, 2000).

2. Exemplar theory and the usage-based model

The usage-based model in grammar has been proposed for many of the same reasons that the exemplar model has been proposed for phonology (and Bybee's work applies to both phonology and grammar). Bybee and others have demonstrated the reality of frequency effects in the representation of morphological and syntactic form, and in processes that lead to the reshaping of form such as analogical levelling in morphological paradigms. Bybee and Scheibman (1999)

investigate the variable phonetic realization of tokens of English *don't* and offer a frequency-based explanation of the probability distribution of the phonetic reduction of *don't*.

On the semantic side, research on polysemy, especially corpus-based approaches, have demonstrated a high degree of variability in the meanings of linguistic expressions (seminal studies include Lindner 1981, Brugman 1988 and Lakoff 1987). This observation has led to a series of studies in cognitive linguistics following a similar method of analysis. A construction (in the broad sense, including individual words or grammatical morphemes) is investigated, ideally via a corpus of attested language use, but also through introspection. A range of meanings is found for the construction, confirming the variability in the use of that construction. A radial category or polysemy network is constructed which represents the semantic relatedness of the functions of the construction in terms of their semantic similarity; this network is usually constructed via an a priori semantic analysis of the functions of the construction. Finally, a prototype or set of core functions of the construction is identified, using criteria such as synchronic token frequency or a diachronic order of uses.

A typical example of this approach is Cuyckens (1995). Cuyckens investigates the range of functions of the Dutch preposition *door* (roughly translated as 'through'). He identifies five distinct functions who semantic structures are represented in diagrams indicating the location or movement of the figure (trajector) relative to the ground (landmark). The five functions are related to one another in terms of the network in Figure 1 below:

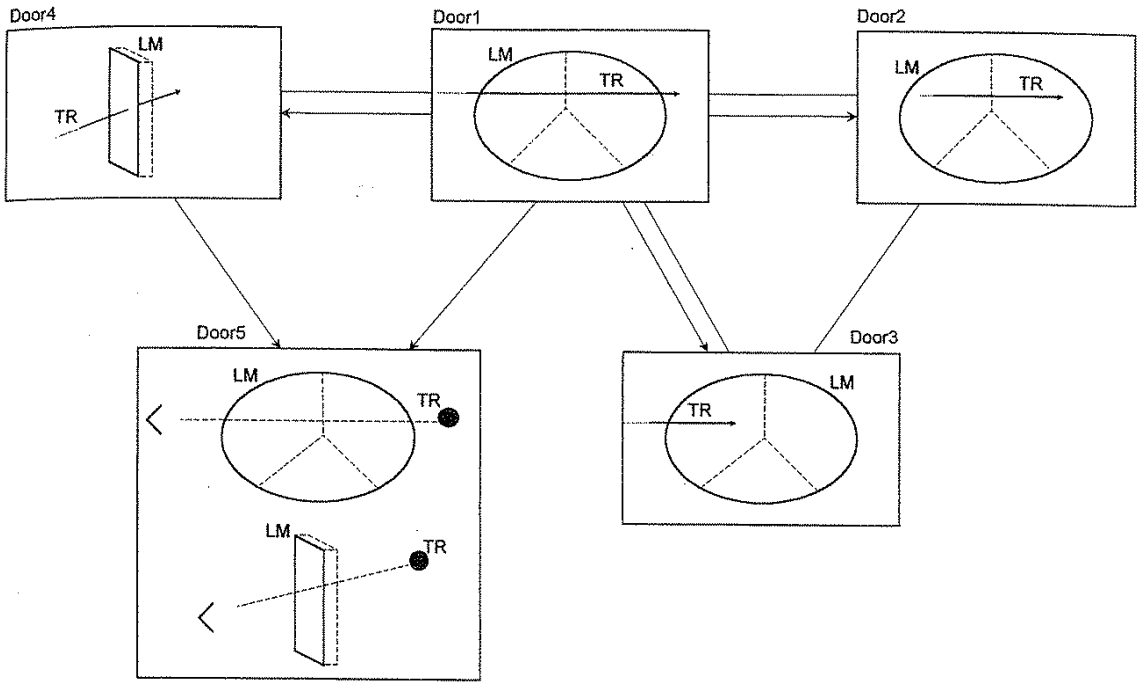

Figure 1. Radial category of schemas for Dutch *door*
(from Cuyckens 1995:197)

The function designed Door1 is the prototype, according to Cuyckens, based on the fact that its semantic properties motivate the extension of *door* to the other spatial scenes represented by Door2-Door5.

This classic type of usage-based analysis of meaning has a number of methodological characteristics. First, it has a comprehension orientation, moving from form to meaning: it represents the range of meanings to which a user has heard that form being applied. Second, it chronicles the number of functions of a form, including a frequency distribution of tokens of the form if the analysis is based on a corpus. Third, it is oriented to the internal structure of a category, that is the network structure of the functions of the form, in particular as they are organized in terms of a prototype and its extensions.

The usage-based model is rather different from the exemplar model as applied to phonology. The usage-based model identifies abstract functions, not actual tokens of use. The exemplar

model in phonology is production-oriented: the variability in phonetic realization is a consequence of the production process, although it has consequences for the listener in terms of his memory of tokens of phonological production. In grammar, we must also examine the forms used for a particular function. This corresponds to what a speaker is doing: she begins with an experience to be verbalized, and the product of the verbalization process is an utterance in a particular grammatical form. When this is done, we find that there is also a high degree of variability, just as in the phonetic realization of a phoneme (see §3.1). I argue that the expression of meaning in linguistic form must be defined as a probability distribution of forms for a particular function. Just as phonetic values are mapped onto a phonetic space, formal "values" (instances of constructions) must also be mapped onto a syntactic space (see §3.1 and §3.2).

Finally, in addition to identifying the internal structure of the category, we must investigate the properties of category boundaries. Category structure involves not just the relationships between tokens in a space (phonetic, syntactic or conceptual); it also involves boundaries which distinguish categories and group together functions that are similar in certain ways (Croft and Cruse 2004, ch. 4). In fact, this helps us to construct the conceptual space in a non-a priori, inductive, empirical linguistic fashion (see §3.3).

3. Exemplar semantics from a production orientation

3.1. Identifying forms used for a function

In exemplar phonology, it is fairly easy to ensure the identity of phonemes: although there are of course problematic cases, one can identify phonological categories given assumptions about how they are to be found across different words. Also, it is fairly easy to collect large numbers of tokens of the same phoneme in order to determine the probability distribution of phonetic realizations of a phoneme in phonetic space.

Doing the same in grammar is far more difficult. First, it is not clear what counts as the same semantic situation, or more precisely, the same experience that is to be verbalized. Second, it is even more difficult to collect large numbers of tokens of the verbalization of the same situation. Technically, every situation is unique, and even if we take a granular view of conceptual space, it is virtually impossible to guarantee that we are looking at the verbalization of the same experience in naturalistic settings. However, we can design similar situations from speaker to speaker, and elicit verbalizations of those situations from multiple speakers (and multiple languages; see §3.2 and §3.3). The same depicted situations are shown to different speakers in identical circumstances, and verbalizations elicited from speakers in identical circumstances, to maximize comparability. Examples of this experimental design are the Pear Stories film (Chafe 1980), in which a film without language was designed and produced; the Bowerman-Pederson spatial pictures, used to elicit the same set of spatial relations across speakers and languages (see Levinson et al. 2003; §3.3 below), and the cutting/breaking video clips used to elicit the same set of events across speakers and languages (Majid et al. 2004).

In this section, I describe some results from a study of the verbalization patterns found in the twenty English Pear Stories narratives found in Chafe (1980; a full description of the study is found in Croft, to appear). The narrative was divided into scenes, using the subchunking of the movie events that emerged from comparing the verbalizations of the twenty speakers. Verbalizations of specific verbs and constructions of various types were tallied, scene by scene. Individual scenes were analzyed separately at first, in order to maximize the identity of the experience being verbalized.

The most general and clearest result is that there is variability in the verbalization of almost every scene. An example of the variability found is given in (1), the verbalizations of scene D5

(for the labeling of scenes, see Croft to appear; the numbers x,y refer to speaker and intonation unit).

(1) Verbalizations of scene D5 from the Pear Stories

1,75	[.45] he when he turns around his hat flies off.
2,65	[1.05 [.55] and uh] it turns out she [.7] from what I could understand she grabbed his hat.
3,20	[.9 [.7] uh] he loses his hat,
6,33	[.6] and his hat flies off,
7,49	{cross}=and she knocks the hat that he's wearing off on the ground,
8,28	[.7 [.1] a--nd] his hat falls off,
10,93	[.5] and apparently he [.9] I think by the breeze,
10,94	. . his hat sort of gets [.7] blown off his head=
11,66	[.5 . . And [.3]] his hat blows off,
11,67	[.55] when they cross,
12,108	[.8] also,
12,109	. . before he fell over,
12,110	[.2] his hat blew off.
12,111	[.25] While he was still looking at the girl.
13,57	and she brushes off this little hat that he has on,
13,58	[.7] and so his hat . . comes o--ff,
14,70	. . lost his hat,
15,62	[.8] and he checks [.3] and his hat flies off also.
17,99	[.35] The little boy {creaky sound} . . that was on the bike,
17,100	had been wearing a hat.
17,101	[1.3 [.55] A--nd [.3]] in the [.55] i--n passing the little girl,
17,102	it had . . fallen off.
18,34	so that his [.6] his hat flies off.
19,57	his hat comes off,
20,25	[.35+ and [.35]] somehow she took his hat.
20,26	. . Not on purpose but [.8] it came off.

This high degree of variability is just like the high degree of variability in the phonetic realization (vocalization) of phonological categories. In exemplar phonology, the variability can be fairly easily mapped onto phonetic space. In an exemplar approach to grammar, we must find a way to map the syntactic variation in language use onto a syntactic space. I suggest that the dimensions of syntactic space be structured in terms of the model of the verbalization process proposed by Chafe (1977a,b) and extended in Croft (2007). Chafe tackles the problem of how a speaker takes an experience which is a unique whole, not structured like the usual propositional semantic representations in language production models, and gives it a structure that can ultimately be verbalized. Chafe proposes three processes. The first, **subchunking**, breaks up the whole experience into chunks that can ultimately be verbalized in a single utterance (or perhaps more specifically, a single clause, leaving aside problems in individuating clauses). The second, **propositionalizing**, takes a chunk, extracts entities from it that are likely to persist across chunks (roughly, the individuals functioning as arguments), leaving the rest of the scene as roughly the predicate (and possibly other semantic components). The third process, **categorizing**, takes the propositionalized entities and categorizes them as belong to types that recur (e.g. categorizing a particular entity in the experience as a hat or flying-off).

The product of these three processes (which do not necessarily proceed in sequence) are the lexical items that categorize the parts of the experience that have been subchunked and propositionalized. In Croft (2007), I argue that grammatical morphemes and constructions serve to, so to speak, "restore" or at least evoke the original unique whole of the experience. Various grammatical inflections and constructions serve to take the semantic types that result from the categorization process, and evoke the particular individual involved in the original experience. I call these processes **particularizing**, and they include situating the particular entity (individual

or event) in space and/or time. Clause level constructions such as argument structure constructions put back together the individuals and the event, indicating who did what to whom. I call these processes **structuring**. Finally, clause linking and reference tracking constructions serve to link together the chunks to evoke the original whole experience; I call these processes **cohering**.

The result is a completely instantiated sentence, a **construct** (to use the Construction Grammr term), with all of its words and constructions assembled. Table 1 presents the dimensions of syntactic space for the verbalizations of scene D5 in accordance with these verbalization processes in tabular form (subchunking and propositionalizing are not described here, or rather only implicitly in the individuals and events verbalized by different speakers):

Table 1. Constructing a syntactic space (= verbalization space; cf. Chafe 1977a,b; Croft 2007)

	Categorizing				**Particularizing** ***Situating***		
	Boy	*Girl*	*Hat*	*Action*	*Hat*	*Action-Time*	*Action-Mod*
1	–	–	hat	fly off	–	Pres	Decl
2	–	–	hat	grab	–	Past	Decl
3	he	–	hat	lose	–	Pres	Decl
6	–	–	hat	fly off	–	Past	Decl
7	–	she	hat	knock off	that he's wearing	Pres	Decl
8	–	–	hat	fall off	–	Pres	Decl
10	–	–	hat	get blown off	–	Pres	apparently
11	–	–	hat	blow off	–	Pres	Decl
12	–	–	hat	blow off	–	Past	Decl
13	–	–	little hat	brush off	that he has on	Pres	Decl
	–	–	hat	come off	–	Pres	Decl
14	–	–	hat	lose	–	Past	Decl
15	–	–	hat	fly off	–	Pres	Decl
17	–	–	it	fall off	–	Pluperfect	Decl
18	–	–	hat	fly off	–	Pres	Decl
19	–	–	hat	come off	–	Pres	Decl
20	–	she	hat	take	–	Past	somehow
	–	–	it	come off	–	Past	Decl

	Structuring			**Cohering** ***Reference Tracking***			***Clause Linkage***
	Sbj	*Obj*	*Obl*	*Boy*	*Girl*	*Hat*	
1	Pat	–	–	–	–	Poss	when he turns around
2	Agt	Pat	–	–	–	Poss	and it turns out
3	Exp	Pat	–	Prn	–	Poss	Ø
6	Pat	–	–	–	–	Poss	and
7	Agt	Pat	Loc	(Prn)	–	Def	and
8	Pat	–	–	–	–	Poss	and
10	Pat	Exp-Part	Force	–	–	Poss	and
11	Pat	–	–	–	–	Poss	and
12	Pat	–	–	–	–	Poss	also, before he fell over,
13	Agt	Pat	–	–	Prn	this	while he was still looking at the girl,
	Pat	–	–	–	–	Poss	and so
14	Exp	Pat	–	Ø	–	Poss	Ø
15	Pat	–	–	–	–	Poss	and…also
17	Pat	–	–	–	–	Prn	and in passing the little girl,
18	Pat	–	–	–	–	Poss	so that
19	Pat	–	–	–	–	Poss	Ø
20	Agt	Pat	–	–	Prn	Poss	and
	Pat	–	–	–	–	Prn	but

This approach also allows the analyst to reduce the extreme variability of the full utterance to the variability of lexical and constructional parts of the utterance. In Croft (to appear), this was done for a variety of words and constructions. In all cases, I found that in comparing similar scenes, the differences in verbalization were not categorical but probabilistic: the same verbs/constructions were used for the different scenes, but in different proportions. Moreover, in a number of cases, the differences in the proportions of verbs/constructions used correspond clearly to differences between otherwise similar scenes.

For example, the Pear film included several scenes that involved a human participant who did not intend to bring about the event; this participant is hence an experiencer or undergoer. Grammatically, the experiencer (undergoer) is encoded either as a subject (2), as a nonsubject (3-4), or is not expressed at all, an existential construction being used instead (5).

(2) 2,67 and then **he** . . crashes into a rock.

(3) 11,68 [1.2 [.25] and [.65]] **his bike** hits into a rock,

(4) 7,53 [.25] and **the pears** all [.45] spill on the ground,

(5) 3,21 a--nd . . **there's a stone** in the way,
3,22 so his bicycle falls over,

Table 2 gives the distribution of these different grammatical encodings of the human participant in the scenes of unintended human events (adapted from Croft to appear, Table 11):

Table 2. The verbalization of unintended human events (Croft to appear, Table 11)

	Exp/Und-Sbj	Other-Sbj	Exist	Other	***Total***
D8. The cyclist falls/bike falls	15	2	–	2	***19***
D7. The cyclist hits a rock/bike hits rock	14	5	3	–	***22***
A4. Drop pears/pears drop	1	2	–	–	***3***
D5. The cyclist loses hat/hat flies off	2	11	–	–	***13***
G4. He is missing a basket/the basket is missing	2	12	5	–	***19***
D9. He spills pears/The pears spill	2	17	–	1	***20***

The scenes with a higher proportion of experiencer/undergoer subjects (D7, D8) contain events which are more likely to be assumed to be under the control of the human participant (the cyclist and his bicycle). The scenes with a lower proportion of experiencer/undergoer subjects (A4, D5, D9, G4) contain events which are more likely to be assumed to be out of control of the human participant. But the difference in verbalization is not categorical: each scene has verbalizations with an experiencer as subject and verbalizations with an experiencer in another grammatical role. It is a difference in the probability distribution of the argument structure constructions that tells us that we are dealing with semantically different scenes.

Likewise, with the second mention of nonhuman entities. The second mention is alternatively verbalized with a simple definite article or a possessive pronoun in English, as seen from these two examples of reference to the pears in scene A7:

(6) 1,16 and he [.3] dumps all **his pears** into the basket,

6,10 and dumps **the pears** into a basket.

Table 3 summarizes the distribution of second mentions of nonhuman referents across the scenes in which they occur (data derived from Croft to appear, Table 6):

Table 3. The verbalization of second mentions of nonhumans

	Def	Poss	Other	***Total***
tree (13 scenes)	44	1	–	***45***
goat (2 scenes)	9	1	1	***11***
ladder (5 scenes)	21	3	–	***24***
pears (6 scenes)	43	13	14	***70***
bicycle/bike (2 scenes)	8	20	–	***28***
hat (2 scenes)	12	23	2	***37***
apron (2 scenes)	–	4	–	***4***

The probability distribution of definite article vs. possessive pronoun reflects alienability, animacy and canonical relationships between the human participants and the objects mentioned. The fewest possessive pronouns are found with the more alienable and more animate objects (the tree and the goat). The most possessive pronouns are found with the more inalienable and less animate objects (the hat and the apron, both closely associated with the human body). Regarding the objects of intemediate alienability, the ladder and the pears are less likely to be owned by the worker whereas the bicycle is more likely to be owned by the cyclist. These differences in canonical ownership relations are reflected in the probability distribution of the verbalizations:

fewer possessive pronouns for the ladder and the pears, and more possessive pronouns for the bicycle.

A final example is the verbalization of events with the verb *see*. In Croft (to appear), I investigated the verbalization of scenes using verbs such as *see* which tend to be the sources of grammatical elements in grammaticalization. The probability distribution and the other verbs used for scenes in which *see* was used by at least one speaker are given in Table 4:

Table 4. The verbalization of see *(from Croft to appear, Table 4)*

	See	***Other Verb***	***Other verbs used***
C3 (cyclist-pears)	4	2	*look at*
E4 (boys-cyclist)	7	–	
F1 (boys-hat)	3	12	*notice, find, come across, run across*
G4	3	15	*notice, discover, realize, look at*

The distribution of verb verbalizations indicate the semantic differences between the events in the respective scenes. Scene G4 clearly represents a cognition event: English *see* can be used as a verb of cognition like *notice* and *realize*, which were used more commonly for this scene. Scenes C3 and E4 indicate a more prototypical use of *see* as a perception verb, alternating with the aspectually distinct activity verb *look at*. In fact, 'see' verbs are frequently etymologically derived from 'look at' verbs, evidence for their semantic similarity (Croft to appear, citing Buck 1949, §15.51: 'see' < 'look at' in many Indo-European languages). Scene F1 has a different distribution of verbs, which do not fit the pattern of semantic origin for 'see' verbs. However, the distribution of verbs for Scene F1 is typical for 'find' verbs and their etymological sources (Croft to appear, citing Buck 1949, §11.32: 'find' < 'see', 'notice' 'come/run across'). In scene F1, the

boys ‘see’ the hat that the cyclist has lost—much more of a finding type situation than a prototypical seeing situation.

This data leads to the following conclusions for the representation of grammatical knowledge. There is evidence for very fine-grained differences among different semantic situations in the linguistic behavior of a set of speakers verbalizing the same scenes. These fine-grained differences can be captured by an exemplar model of meaning. The evidence, however, is not in the form of a uniform one-to-one relationship across speakers between the semantic exemplar and a precise grammatical description, but rather as a probability distribution of different words or constructions used by different speakers for a particular scene. This evidence implies that the best way to relate the semantic exemplars to grammatical form is by a probability distribution of words and/or constructions used for each semantic exemplar. (This is incidentally a different and more reliable frequency measure for the form-function relationship than the token frequency of different functions of a single grammatical form. The token frequency of different functions of a form does not take into account the overall token frequency of each function/situation type.)

Thus, we arrive at a model of grammar in which the form-function mapping is quite complex. Semantic exemplars are mapped onto a probability distribution of linguistic forms that have been used to verbalize the semantic exemplar. The different linguistic forms are mapped onto a syntactic space whose dimensions are structured according to the processes involved in verbalizing experience. The semantic exemplars are themselves mapped onto a conceptual space which organizes them.

3.2. Syntactic space in a typological analysis

It is difficult to carry out such semantically fine-grained studies because of the experimental controls required; the studies that offer evidence for an exemplar-based approach are quite recent. However, typologists have performed the same kind studies across languages for years, albeit in a far more coarse-grained fashion: select a set of related meanings that represent coarse-grained exemplars in conceptual space, and analyze the variation in the verbalizations of those exemplars across languages. Such research has often revealed a grammatical continuum in verbalization, which I proposed to represent in a syntactic space in Croft (2001). The prime example given there is grammatical voice. Relatively superificial analyses suggested that there are three distinct voice types across languages: active/direct, passive and inverse. The active and passive types are illustrated in English in (7)-(8), and the direct/inverse types are illustrated by Cree in (9)-(10) (Wolfart & Carroll 1981:69):

(7) They took the boy to school.

(8) The boy was tak**en** to school (**by** his parents).

(9) ni- wāpam **-ā** -wak
1- see -**DIR** -3PL
'I see them'

(10) ni- wāpam **-ikw** -ak
1- see -**INV** -3PL
'They see me.'

In the active or direct voice, the A (roughly, agentive) participant is coded like a subject, the P (patientive) participant is coded like an object. In the passive, P is coded like a subject, and A like an oblique (neither subject nor object). The inverse differs from the passive: although P is coded like a subject, A is coded like an object.

However, there are many voice constructions which are intermediate between these "types" in every possible way. Two examples suffice here; see Croft (2001, ch. 8) for more examples. The Welsh Pronominal Passive codes P like an object but A like an oblique (Comrie 1977:55):

(12) fe**'i** lladdwyd **gan** ddraig
PRT**'OBJ** killed.**PASS** **by** dragon
'He was killed by a dragon.'

The Arizona Tewa Passive (Kroskrity 1985:313)/Inverse (Klaiman 1991) codes the P like a subject and A like an oblique in terms of case marking, but codes the A-P relation with a special agreement form:

(14) ụ kʰóto hẹ'i sen **-di** **wó:-** mégi
you bracelet that man **-OBL** **2/3.PASS-** give
'You were given a bracelet by that man.'

My observations were not new in the typological literature, as the following quotations show:

> The analysis of the various constructions referred to in the literature as PASSIVE leads to the conclusion that there is not even one single property which all these constructions have in common (Siewierska 1985:1)

> I know of no structural features which can define inverse constructions and distinguish them from passives (Thompson 1994:61)

> Passives form a continuum with active sentences (Shibatani 1985:821)

In Croft (2001), I plotted the different voice constructions found across languages in a syntactic space based on the coding of the A and P participants. The syntactic space is

reproduced as Figure 2 here (Croft 2001:313, Fig. 8.13; this is essentially a two-dimensional representation of the dimensions of the structuring verbalization process found in Table 1):

	P CODING			
	SBJ-LIKE	SPEC	DIR	OBJ-LIKE
A CODING				
SBJ-LIKE				ACTIVE/DIRECT
				Acehnese E
			Dyirbal SE	
				Maasai "I"
SPEC		Seko Padang "I"		
	Karo Batak P			
	Kapampangan GF			**Chukchi "I"**
	Cebuano GF			
DIR	**Cree I_{prn}**		Guaraní "I"	
		Yurok "I"		
		Arizona Tewa "I"/"P"		
	Shilluk P		Tangut I	
		Upriver Halkomelem P		
	Cree I_{obv}			
	Indonesian P			
	Pukapukan E			
	Pukapukan P			
	Bambara "P"		Bella Coola P	
OBL-LIKE	**English P**	Spanish RP		Russian IP
				Welsh IP
PROHIBITED				**Menomini P**
	Lithuanian P	**Finnish ID**		**Maasai P**

A - Active
P - Passive
IP - Impersonal Passive
RP - Reflexive Passive
ID - Indefinite
I - Inverse
E - Ergative
SE - Split Ergative
GF - Philippine Goal Focus

boldface: verb form distinct from Active/Direct verb form

Scaling (A top to bottom, P left to right):
A case: sbj < erg < dir < obl < prohibited
A agr: sbj < nonsbj < special < none < prohibited
P case: sbj < dir < obj
P agr: sbj < special < obj/none

Figure 2. Two-dimensional spatial model of the syntactic space for voice constructions

In the same work, I argue that the syntactic space can be mapped onto a conceptual space governed by the degree of salience (down to complete absence) of the A and P arguments, which in turn is embedded in a conceptual space that includes events with only one argument (see Figure 3; Croft 2001:317, Fig. 8.16):

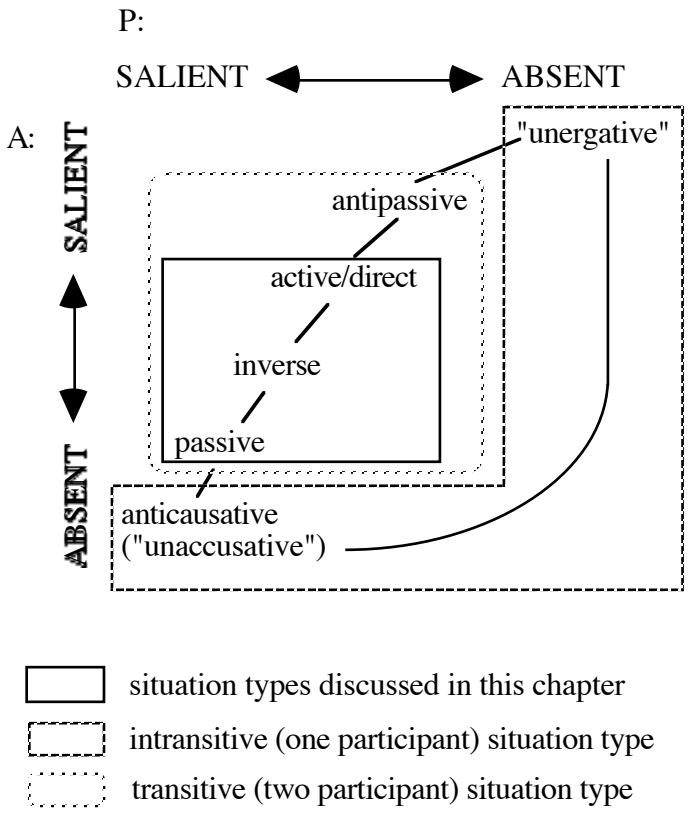

Figure 3. The conceptual space for voice and transitivity

The typological analysis in Croft (2001) summarized here suggests that there is in fact a close relationship between the structure of the syntactic space and the structure of the conceptual space. This conclusion, if generally correct—and that is the general thrust of the results of

typology—significantly constrains the mapping between the semantic exemplars in the conceptual space and the syntactic exemplars (particular constructs) in syntactic space.

3.3. Using category boundaries to construct a conceptual space

In §3.1 and §3.2, we noted that a linguist practicing exemplar semantics, or more generally exemplar grammar, faces major empirical problems in determining the probability distribution of grammatical forms for particular semantic exemplars. The ideal solution for this problem is the experimental verbalization paradigm. Another solution is the traditional typological analysis, which uses a more coarse-grained set of semantic exemplars but uses cross-linguistic variation as another approach to the probability distribution of grammatical structures for coarse-grained semantic exemplars.

A second problem is determining the structure of conceptual space in the first place. The most common method in cognitive linguistics is an a priori one: to use one's intuitive semantic analysis to construct the network of exemplars. However, if we construct the conceptual space a priori, we may overlook semantic dimensions that are relevant to the analysis. More generally, we may put too much or too little weight on certain semantic distinctions over others.

In the past decade or so, an empirical method has been used in typology to allow the structure of the conceptual space to emerge from the categorizations used by speakers of languages. This is the semantic map model (see Haspelmath 2003 for a survey). The basic principle behind the semantic map model is that if two functions are expressed by the same form, then they have been judged as similar by at least some speakers. If there is an underlying conceptual space, then we should be able to construct it by simultaneously comparing the categorizations by speakers of each language. The semantic map model represents a conceptual space as a graph structure of linked semantic exemplars. An example is Haspelmath's semantic map of indefinite pronouns

(Hasplemath 1997:64, Fig. 4.4; see Haspelmath 1997 for definitions and examples of the semantic exemplars).

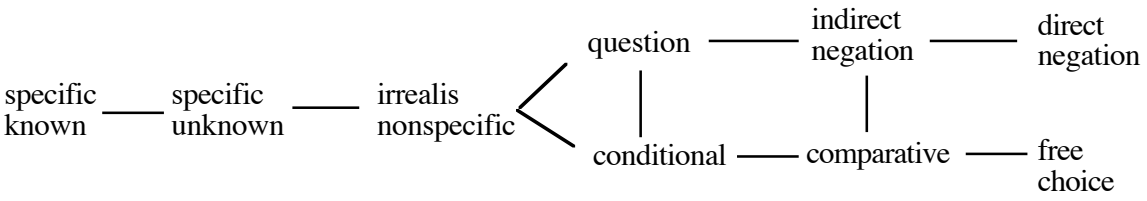

Figure 4. Conceptual space for indefinite pronoun functions

Particular language categories are mapped onto the conceptual space, as in Figure 5 for the Romanian indefinite pronouns:

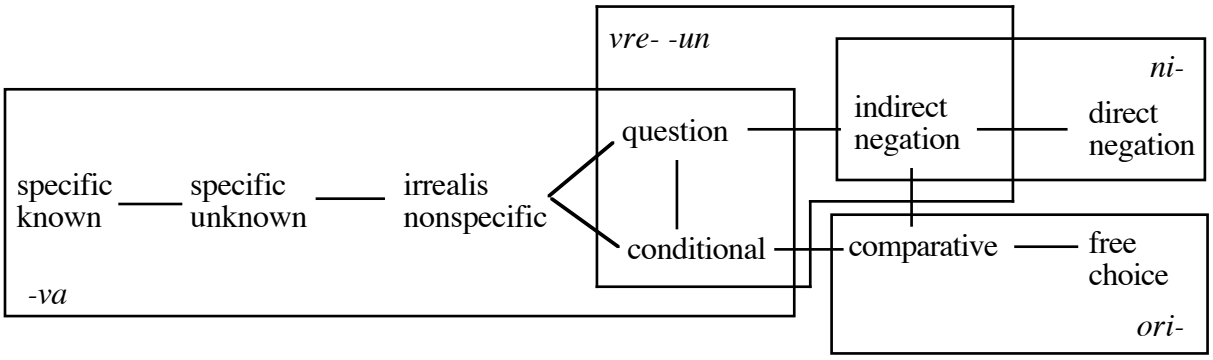

Figure 5: Semantic maps of Romanian indefinite pronouns

The similarity principle is manifested by the constraint that any language-specific grammatical category must include a connected subgraph of the conceptual space (i.e., no discontinuous grammatical categories).

This approach to the construction of conceptual space has the advantage of being based on the empirical facts of speaker's linguistic behavior, rather than linguist's a priori assumptions (which are often correct, but not always so). Another important point about this method is that it demonstrates that category boundaries are linguistically and semantically significant. In

cognitive linguistics, category boundaries are often ignored in favor of the internal structure of the category, namely its prototype and extensions. Yet boundaries are as important for categories as their internal structure. Boundaries provide similarity information about semantic exemplars, and hence are sensitive to the structure of conceptual space. Boundaries therefore provide evidence for linguists to uncover the structure of conceptual space.

Nevertheless, the semantic map method has certain practical problems which prevent it from being used for many crosslinguistic studies, especially those involving a fine-grained set of semantic exemplars (Croft and Poole 2008). There is no measure of the goodness of fit of the model of conceptual space (the graph structure) to the crosslinguistic data. In practice, a perfect fit is expected, but this is impractical given the high variability of linguistic data. There is also no interpretation of the spatial dimensions of a semantic map; only the graph structure is relevant. Semantic maps are constructed by hand, and therefore they are constructible for only a small number of semantic exemplars. Finally, there is no mathematically well-understood or computationally tractable technique for constructing semantic maps.

These practical problems can be overcome by using multidimensional scaling (Croft and Poole 2008). Multidimensional scaling (MDS) is a mathematically well-founded technique for constructing spatial models of similarity data. In a spatial model, (dis)similarity is represented as Euclidean distance. Hence the spatial dimensions of the model are interpretable. There are fitness measures which allow the user to determine the number of dimensions for the best-fitting spatial model of the data. The method is computationally tractable, and Poole (2000, 2005) has devised a nonparametric unfolding algorithm, Optimal Classification, which allows the user to analyze categorization data (such as the categorization of scenes by grammatical constructions) directly. In an MDS spatial model, grammatical categories are represented as linear bisections (cutting

lines) of the conceptual space, separating the functions expressed by the grammatical category from the functions that are not. The cutting lines for the Romanian indefinite pronouns are given in Figure 6 for a spatial model of indefinite pronoun functions (from Croft and Poole 2008):

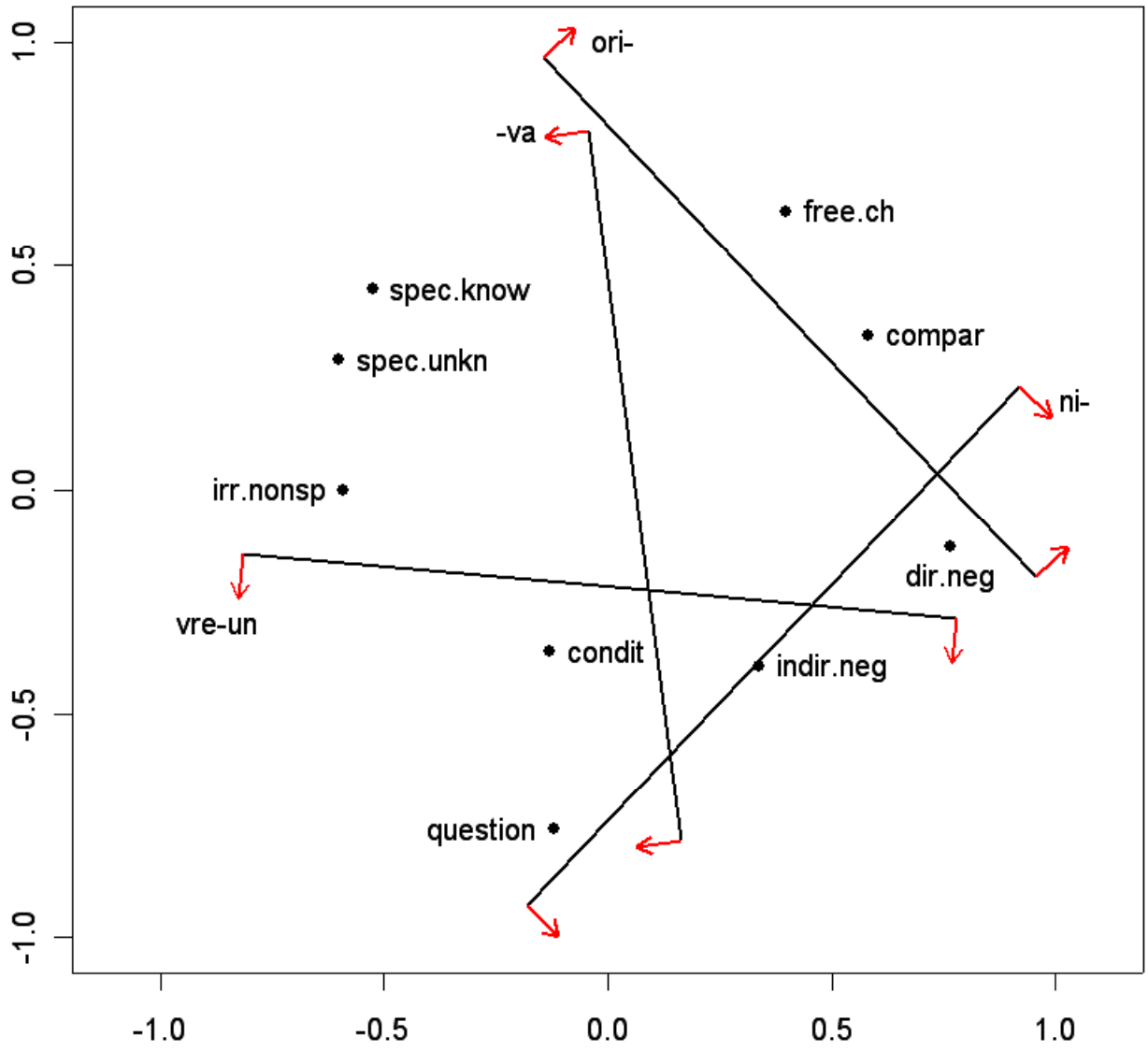

Figure 6: Cutting lines for Romanian indefinite pronouns

MDS and its application to linguistic analysis are described in detail in Croft and Poole (2008). Here I will show its application to one of the fine-grained experimental verbalization

studies mentioned in §3.1, Levinson et al.'s (2003) study of spatial adpositions using the Bowerman-Pederson spatial relations pictures (see also Croft, in prep.). Bowerman and Pederson designed a set of 71 pictures depicting a variety of spatial relations. Each picture represents a semantic exemplar in our terms. Levinson et al. (2003) collected verbalizations of the spatial relations for nine languages (Tiriyó, Trumai, Yukatek, Basque, Dutch, Lao, Ewe, Lavukaleve and Yélîdnye).

Levinson et al. performed an MDS analysis using a dissimilarity algorithm, which requires converting the original data format into a dissimilarity matrix. Unfortunately this compresses the similarity measurements and led to a noisy MDS spatial model. Poole and I used the Optimal Classification algorithm on Levinson et al.'s data (Poole and I are grateful to Sérgio Meira for providing us with the original datasets and MDS analysis from the Levinson et al. study). The fitness measures indicate that a two-dimensional model is best (Classification = percent correct classification; APRE = aggregate proportion reduction of error [see Croft and Poole 2008 for discussion]; adding a third dimension does not substantially improve the fit).

(18)

Dimensions	Classification	APRE
1	94.1	.300
2	95.8	.501
3	97.1	.661

The two-dimensional spatial model is displayed in Figure 7:

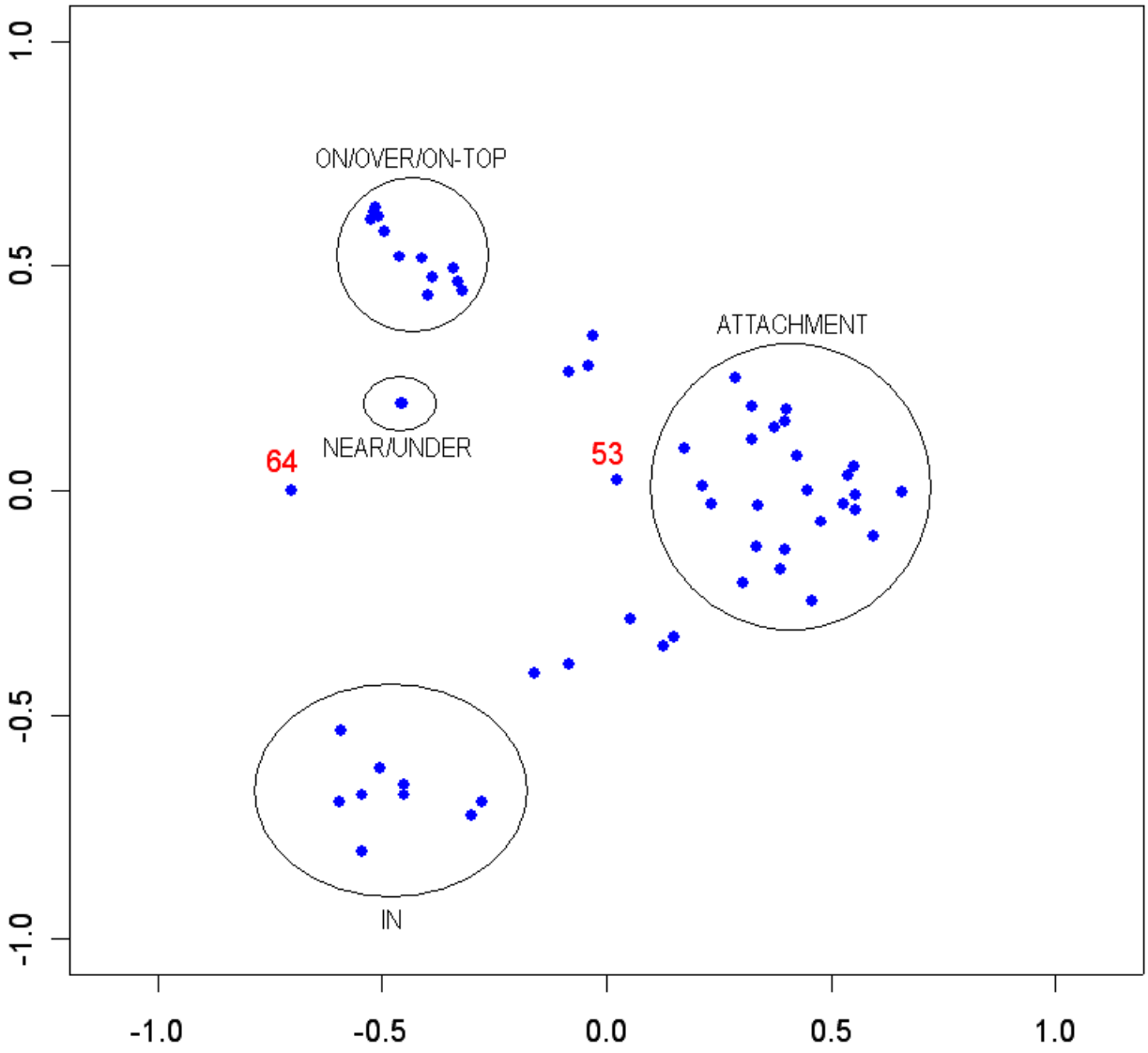

Figure 7: Spatial model of adpositions by unfolding

The Optimal Classification spatial model of the adposition data produces a model of the conceptual space which is intuitively semantically coherent. The clusters indicate regions of conceptual space that represent a topside-oriented spatial relation ('on/over/on top'), an attachment relation, a containment relation, and some sort of a neighborhood ('near/under') relation. The intermediate pictures between the large clusters are semantically intermediate, as expected. The group between the containment and attachment clusters includes five pictures (18,

30, 39, 51, 62). All but picture 51 involve a Figure 'through' a round opening in the Ground—in other words, partial containment and partial attachment. The group between the attachment and topside-oriented clusters includes four pictures (3, 7, 11, 23). All but picture 11 involve some type of surface attachment—in other words, both surface contact/support and attachment.

The MDS analysis of Levinson et al.'s data demonstrate both the practical and theoretical possibility of an exemplar approach to grammar. From a practical point of view, complex and fine-grained linguistic data can be analyzed in order to come up with a representation that respects the exemplar model. From a theoretical point of view, an exemplar approach to the language data leads to a coherent model of conceptual space that is part of the exemplar theory of the relationship between meaning and grammatical form.

4. Conclusions

In this paper, I have sketched an exemplar theory of meaning and grammar, supported by recent empirical studies in verbalization, within and across languages. Exemplars of particular situations that have been verbalized in the experience of the speaker are stored in the speaker's memory. The semantic exemplars are organized by their relations to each other in conceptual space. Semantic exemplars are mapped onto a probability distribution of linguistic forms that have been used to verbalize the semantic exemplar. That is, the probability distribution emerges from the registering of actual tokens of those linguistic forms used for the situation type (semantic exemplar). The different linguistic forms are themselves mapped onto a syntactic space whose dimensions are structured according to the processes involved in verbalizing experience. The syntactic space iconically reflects the structure of conceptual space to a considerable extent, thereby constraining the form-meaning mapping.

The exemplar approach to meaning and grammar is radically nonreductionist (Croft 2001:47-48). Each situation/scene as a whole is a primitive element in the representation, i.e. a point in conceptual space. To put it another way, each semantic frame is a semantic primitive. Likewise, each construct is a primitive element in the representation, a point in syntactic space. The formation of more general semantic and syntactic structures is a process of abstraction over actual exemplars of situations and utterances, guided by the similarity relations between situations and between the forms of utterances. While much remains to be explored in such a model, I believe it is the best approach for the analysis of linguistic form and meaning available at present.

References

Brugman, Claudia Marlea. 1988. *The Story of* Over*: Polysemy, Semantics, and the Structure of the Lexicon*. New York: Garland Publishing.

Buck, Carl Darling. 1949. A dictionary of selected synonyms in the principal Indo-European languages. Chicago: University of Chicago Press.

Bybee, Joan L. 1985. *Morphology: A study into the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, Joan L. 2001. *Phonology and language use*. Cambridge: Cambridge Univesity Press.

Bybee, Joan L. & Joanne Scheibman. 1999. The effect of usage on degrees of consitutency: the reduction of *don't* in English. *Linguistics* 37.575-96.

Chafe, Wallace. 1977a. Creativity in verbalization and its implications for the nature of stored knowledge. *Discourse production and comprehension*, ed. Roy Freedle, 41-55. Norwood, New Jersey: Ablex.

Chafe, Wallace. 1977b. The recall and verbalization of past experience. *Current issues in linguistic theory*, ed. Peter Cole, 215-46. Bloomington: Indiana University Press.

Chafe, Wallace (ed.). 1980. *The Pear Stories*. New York: Ablex.

Comrie, Bernard. 1977. In defense of spontaneous demotion: the impersonal passive. *Grammatical Relations*. (Syntax and Semantics, Vol. 8.), ed. Peter Cole and Jerrold M. Sadock, 47-58. New York: Academic Press.

Croft, William. 2000. *Explaining language change: an evolutionary approach*. Harlow, Essex: Longman.

Croft, William. 2001. *Radical Construction Grammar: syntactic theory in typological perspective*. Oxford: Oxford University Press.

Croft, William. 2007. The origins of grammar in the verbalization of experience. *Cognitive Linguistics*.

Croft, William. to appear. The origins of grammaticalization in the verbalization of experience. *Linguistics*.

Croft, William. In preparation. Relativity, linguistic variation and language universals. To appear in a volume edited by Maarten Lemmens.

Croft, William and D. Alan Cruse. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press.

Croft, William and Keith T. Poole. 2008. Inferring universals from grammatical variation: multidimensional scaling for typological analysis. *Theoretical Linguistics*.

Cuyckens, Hubert. 1995. Family resemblance in the Dutch spatial prepositions *door* and *langs*. *Cognitive Linguistics* 6.183-207.

Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Oxford University Press.

Haspelmath, Martin. 2003. The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. *The new psychology of language*, *vol. 2*, ed. Michael Tomasello, 211-42. Mahwah, N. J.: Lawrence Erlbaum Associates.

Klaiman, Miriam. H. 1991. *Grammatical voice*. Cambridge: Cambridge University Press.

Kroskrity, Paul V. 1985. A holistic understanding of Arizona Tewa passives. *Language* 61.306-28.

Lakoff, George. 1987. *Women, fire and dangerous things: what categories reveal about the mind*. Chicago: University of Chicago Press.

Langacker, Ronald W. 1988. A usage-based model. *Topics in cognitive linguistics*, ed. Brygida Rudzka-Ostyn, 127-161. Amsterdam: John Benjamins.

Langacker, Ronald W. 2000. A dynamic usage-based model. *Usage-based models of language*, ed. Michael Barlow and Suzanne Kemmer, 1-63. Stanford: Center for the Study of Language and Information.

Levinson, Stephen C., Sérgio Meira, and the Language and Cognition Group. 2003. 'Natural concepts' in the spatial topological domain—adpositional meanings in crosslinguistic perspective: an exercise in semantic typology. *Language* 79.485-516.

Lindner, Susan. 1981. A lexico-semantic analysis of the English verb particle constructions with *out* and *up*. Ph.D. dissertation, University of California, San Diego.•

Majid, Asifa, Miriam van Staden, James S. Boster, and Melissa Bowerman. 2004. Event categorization: a cross-linguistic perspective. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, 885-890.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: word frequency, lenition and contrast. *Frequency and emergence in grammar*, ed. Joan L. Bybee & Paul Hopper, 137-57. Amsterdam: John Benjamins.

Pierrehumbert, Janet B. 2003. Probabilistic phonology: discrimination and robustness. *Probabilistic linguistics*, ed. Rens Bod, Jennifer Hay and Stefanie Jannedy, 177-228. Cambridge, Mass.: MIT Press.

Poole, Keith T. 2000. Non-parametric unfolding of binary choice data. *Political Analysis* 8(3).211-237.

Poole, Keith T. 2005. *Spatial models of parliamentary voting*. Cambridge: Cambridge University Press.

Shibatani, Masayoshi. 1985. Passives and related constructions: a prototype analysis. *Language* 61.821-48.

Siewierska, Anna. 1985. *The Passive: A Comparative Linguistic Analysis*. London: Croom Helm.

Thompson, Chad. 1994. Passive and inverse constructions. *Voice and inversion*, ed. Talmy Givón, 47-63. Amsterdam: John Benjamins.

Wolfart, H. Christoph & Janet F. Carroll 1981. *Meet Cree: A guide to the Cree language* (2nd edition). Lincoln: University of Nebraska Press.