Linguistic Issues in Language Technology – **LiLT**
Submitted, October 2011

Variety, Idiosyncrasy, and Complexity in Language and Language Technologies

Lori Levin

Published by CSLI Publications

LiLT volume 6, issue 10 October 2011

Variety, Idiosyncrasy, and Complexity in Language and Language Technologies

Lori Levin, *Carnegie Mellon University*

1 Introduction

Until 1996 the annual meeting of the Association for Computational Linguistics did not have parallel sessions. The field was small, and most participants knew at least a little bit about all of the sub-fields. It was common practice for students in computational linguistics to take courses in linguistics and computer science. The author of this paper taught in the former Computational Linguistics Program jointly run by Carnegie Mellon University the University of Pittsburgh from 1986 to 1996. Students in that program took two semesters of syntax, one semester of semantics, and one semester of pragmatics. They wrote original qualifying papers in linguistics and also did core qualifying courses in computer science. This paper is not advocating that current language technologies students become equally well versed in linguistics and computer science. The field has now reached a point where no one can know all parts of it well. The paper is, however, claiming that the knowledge of linguists and computer scientists in the field has become disjoint, and it suggests some ways to re-introduce linguistics to computer scientists.

This paper addresses three issues in language technologies. For each issue, the paper recommends an area of linguistics that is easily accessible to computer scientists and provides some examples that may be thought-provoking. The first issue is linguistic diversity, which is ad-

LiLT Volume 6, Issue 10, October 2011.
Variety, Idiosyncrasy, and Complexity in Language and Language Technologies.

dressed by language *typology*. Typology provides an insightful view of the syntax and semantics of word order, as presented in Section 2.2. The second issue is the *long tail* of sparse phenomena. Section 3.3 uses *Construction Grammar* as a framework for addressing the details of definiteness and modality. Finally, Section 4 addresses how to make *error analysis* fun. It moves beyond monoclausal sentences and revives some rules from 1970s style transformational grammar as a fun way to analyze complex sentences.

Coverage of human language is daunting. There are more than 6000 languages in the world, and each one has a long tail of sparse phenomena. To make matters worse, many linguistic theories are detailed and inaccessible without years of study. The talent of linguists is to know things about languages they don't speak, or in the words of Annie Zaenen, one of the editors of LiLT, to speak the non-terminals of many languages. This paper recommends three areas of linguistics (typology, Construction Grammar, and old-fashioned transformational grammar) that are critically important to language technologies and are approachable to computer scientists without years of study.

2 Linguistic diversity

2.1 What is a language-independent system?

Bender (2009) points out that the phrase *language independent* used to have a different meaning with respect to language technologies. A technology such as a grammar formalism (LFG, CCG, TAG, HPSG) could claim to be language independent if it was designed to handle what was currently known about human language. More recently, researchers can claim that their algorithms are language independent if no adaptation time is needed to run on a new language. These systems are language independent in the new sense precisely because they are not language independent in the old sense; they do not encode knowledge about human language.

Language independence in the old sense included hurdles that had to be jumped in order to earn the title. The hurdles consisted of challenging phenomena from many languages, such as cross-serial dependencies in Dutch (Bresnan et al., 1982). Some modern competitions like Morphochallenge (Kurimo et al., 2009) preserve the old sense of language independence by providing competition on a range of typologically diverse languages. However, other competitions, including machine translation competitions in the United States, have still not covered some large classes of languages, such as Austronesian (including languages of the Philippines and Indonesia), Bantu (Swahili and related languages), Turkic (a large swath of languages across Central Asia), Japanese, and Korean.

Languages that are not likely to be relevant to the global economy (e.g., indigenous languages of the Americas, Africa, and Asia and some less-spoken languages in Europe) have been even more neglected, and current computational models are even less prepared to handle them. Many of these languages are endangered and are in the process of revitalization. Partial speakers of these languages need lexical and grammatical support for word processing and text messaging. Fluent speakers also need language technologies so that they can continue to use their languages instead of abandoning them in favor of languages that are well supported computationally.

Language typology, the study of the classification of languages, is a useful starting place for language technologists to learn about linguistic diversity. Early computational linguisics was not greatly influenced by typology, although Greenberg's implicational universals about word order (Greenberg, 1963) were well known to many researchers. It was more fashionable for computational linguists to look to generative syntactic theory for language universals. However, the universals were often stated in a theory-specific way and were only accessible to people who invested time in studying the theory. The literature on language typology is much more accessible. *The World Atlas of Langauge Structures* (Haspelmath et al., 2005), for example, has 142 chapters, which are short and self-contained, along with a searchable database and maps with live links.

2.2 A typological perspective on the syntax and semantics of word order

It is generally known that languages differ in basic word order. Subjects may come before or after verbs; adjectives before or after nouns. Most people also know that some languages allow for more variation in word order than others. However, typologists look at word order in terms of both form and function. It is important to distinguish form (in this case, word order) from function (communicating who did what to whom) because the same form can be used for different functions and different forms can be used for the same function, as shown in the examples that follow.

The typological study of word order is significant for language technologies. Word order is not always the primary indicator of grammtical relations and semantic roles (who did what to whom), as it is in English. The importance of other mechanisms such as case marking and agreement should not be underestimated in any multilingual task such as machine translation or cross-lingual question answering. Furthermore, word order is primarily used to express information structure (old and new information) in many languages and should not be mistaken as an indicator of grammtical relations.

Word order is a form whose function is to communicate grammatical relations like Subject and Object, which in turn indicate semantic roles like agent and patient. For example in *The company interviewed the candidate*, the word order tells us that *the company* is the Subject and *the candidate* is the Object. The fact that the verb is in active voice tells us that the Subject (*the company*) is the interviewer and the Object (*the candidate*) is the interviewee. The two-step process from word order to grammatical relations to semantic roles is based on Lexical Functional Grammar (Bresnan, 2000, Dalrymple, 2001). The first step, from word order (or some other formal property of sentences) to grammatical relations, is called *grammatical encoding*. The second step, from grammatical relations to semantic roles, is called *lexical mapping*. The following discussion concerns grammatical encoding.

Many languages use word order for grammatical encoding, but many also rely on morphological processes like case marking and agreement. Languages that have case marking may have fewer restrictions on the order of Subject, Object, and verb. Although the default word order in Russian is SVO, the three Russian sentences in Example (1) can all mean *Tanya killed Masha*. The endings *-a* and *-u* on *Tanya* and *Mašu* indicate that they are Subject and Object even when they are not in the default order.[1]

(1) a. Tanja ubila Mašu.
Tanya.NOM kill Masha.ACC
Tanya killed Masha

b. Tanja Mašu ubila.
Tanya.NOM Masha.ACC kill
Tanya killed Masha

c. Mašu ubila Tanja.
Masha.ACC kill Tanya.NOM
Tanya killed Masha

Chinese and English, two very large languages that have been prominently featured in recent language technologies research, have little or no case marking. Arabic, another prominently featured language, has case marking in the written language but not in the spoken varieties. How important is it for language technologists to understand case?

[1] Sentences provided by Alicia Tribble.

The World Atlas of Language Structures, Chapters 49 and 50 (Iggesen, 2005), considers a sample of 261 languages. One hundred of those languages have no case marking on common nouns (Chapter 49). (English uses case on pronouns e.g., *I* vs. *me*, but not on common nouns.) Eighty-one languages from the sample have no case marking at all (Chapter 50). The remaining 180 languages in the sample have case marking. Major languages with case marking include Russian, German, Japanese, Korean, Hungarian, Finnish, Turkish, Greek, and Hindi.

Agreement is another mechanism for grammtical encoding. Agreement can hold between any kind of head and a dependent: a verb and a noun, a noun and an adjective, etc. The head and dependent may agree in features like person, number, and gender. Agreement between a verb and a Subject and/or Object is relevant to the issue of grammatical encoding. Example (2) from Quechua[2] shows a verb that agrees with both Subject and Object. As in many languages that use agreement for grammatical encoding, the Subject and Object noun phrases can be omitted.

(2) maqa-ma-n
hit-1sg.obj-3sg.subj
He hit me

MacWhinney (2004) discusses sentences like *The cat chase the dogs*, which are not well-formed in Standard English, but illustrate something important about the way English works. The verb *chase* without an *-s* suffix indicates that the Subject should be plural as in the well-formed sentence *The cats chase the dogs*. However, the word order indicates that the singular noun phrase *the cat* is the Subject. If you show this sentence to native English speakers and ask them whether *the cat* or *the dogs* is the Subject of *chase*, they will almost always choose *the cat*. However, Italian speakers will make a different choice. Example (3) poses the same problem as *The cat chase the dogs*. The suffix *-ono* on the verb indicates that the Subject is plural, but the singular noun phrase *il gatto* (the cat) is in the normal Subject position. Italian speakers may choose *i cani* (the dogs) as the Subject of the sentence. In other words, the sentence below means *The dogs chase the cat*. The point here is not that Italian has freer word order than English, but that English speakers will rely on word order for grammatical encod-

[2] Payne (1997): page 135, number 8c with extra glosses verified by Patrick Littell. Varieties of Quechua are spoken by around 10 million people in Peru, Ecuador, Bolivia, Chile, and Argentina (*Ethnologue*, `http://www.ethnologue.com/show_language.asp?code=que`, accessed on December 20, 2010).

ing even if there is evidence, such as verb agreement, to the contrary. Because of this, English speaking researchers may not understand that word order is not the most significant factor for grammatical encoding in many other languages.

(3) Il gatto inseguono i cani.
the-SG cat-SG chase-PL.SUBJ the-PL dog-PL
The dogs chase the cat

The examples presented so far have shown that the communicative function of grammatical encoding can be expressed by different forms including word order, case marking, and agreement. Conversely, the same form (the order of Subject, Object, and verb) can be used to accomplish functions other than grammatical encoding. In many languages changes in word order reflect changes in old and new information.

The effects of old and new information on syntax are well documented (Ward and Birner, 2004, Sgall et al., 1986). Many textbooks (Comrie, 1981, Payne, 1997) explain old and new information with simple question and answer pairs such as *Who killed Masha? Tanya killed Masha.* As soon as this question, *Who killed Masha?* is uttered, *killed Masha* becomes old information, known to both the speaker and hearer. In the answer to the question, the agent of *kill* will be new information. In Russian, the old information is likely to come earlier in the sentence and the new information is likely to be at the end of the sentence. So although the three sentences in Example (1) above are all acceptable in Russian, *Mašu ubila Tanya* would be most appropriate in this context. On the other hand, *Tanya ubila Mašu* would be a good answer to the question *Who did Tanya kill?*, where *Tanya* is the old information. (See Comrie (1981) page 63 for similar examples in Hungarian.)

English also has means for expressing old and new information, but not by altering the basic word order of Subject, verb, and Object. English can use intonation or special constructions such as clefts. Capital letters in the examples below indicate stress (pitch and amplitude). The pound sign in the examples indicates an incoherent discourse. Read the sentences aloud in order to hear the difference between the appropriate and inappropriate sequences.

(4) a. Who killed Masha?
TANYA killed Masha.

b. Who killed Masha?
#Tanya killed MASHA.

c. Who did Tanya kill?
Tanya killed MASHA.

d. Who did Tanya kill?
#TANYA killed Masha.

(5) a. Who killed Masha?
It was Tanya that killed Masha.

b. Who killed Masha?
#It was Masha that Tanya killed.

c. Who did Tanya kill?
It was Masha that Tanya killed.

d. Who did Tanya kill?
#It was Tanya that killed Masha.

Figure 1 summarizes the relation between form (word order, case marking, agreement, and special constructions) and function (semantic roles, grammatical relations, and pragmatic roles) in the languages discussed in this section. Languages can use basic word order, case marking, and agreement as mechanisms for expressing grammatical relations, but basic word order can also be used for indicating pragmatic roles such as old and new information. Figure 1 may be thought-provoking for translation models that incorporate distortion models because the distortion models do not explicitly take function into account. We may wonder whether the models have implicitly adjusted to the communicative functions of word order (grammatical encoding or information structure) or whether they are simply missing something important.

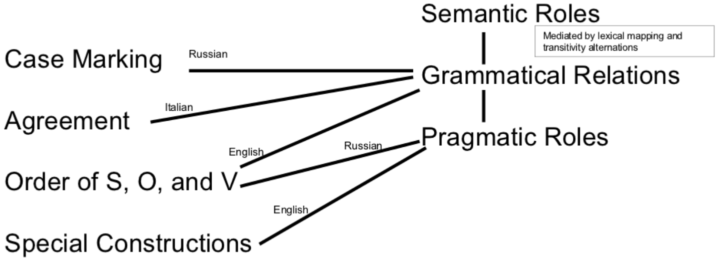

FIGURE 1 **Typology of Grammatical Encoding**

3 The long tail of sparse phenomena

3.1 Are you happy with 80 percent?

The previous section of this paper showed that there is variation in the form-function mapping across languages in the expression of grammatical encoding and information structure. This section looks at form and function from another perspective — the long tail of sparse phenomena, focusing on Construction Grammar and its treatment of idiosyncratic form-function mappings.

The paper by Hajičová in this volume discusses the distinction between core and peripheral phenomena in human languages. Core phenomena include things like the basic order of Subject, Object, and verb. Peripheral phenomena include things like special constructions for exclamations (*What a nice dress!*) or correlation (*The more I read the more I learn*). From a different perspective, the papers by Church and Steedman in this volume refer to a distribution of linguistic phenomena with a large body of frequent phenomena and a long tail of rare phenomena. Although each item in the tail is infrequent, the cumulative mass of the tail is too large to be ignored.

There was a time when the first 80 percent (the core) was considered to be uninteresting because it underdetermined the solution; there were too many ways to achieve it, and comparing systems on how well they handled the core would not reveal which method was truly the best at handling the complexity of human language. The tail of rare phenomena was what distinguished novel work from mundane work. Researchers were expected to know an inventory of universals and peculiarities of language and could try to outdo each other in handling them. It is surprising now that we find ourselves in a field where the inventory has been forgotten and 80 percent is considered to be an excellent score in many classification tasks.

It should be noted, before continuing, that the old days were not always the good old days. In the 1980s papers could be about about toy systems that were not tested on unrestricted or unseen input. Furthermore, the commonly known inventory of universals (e.g., island constraints on filler gap constructions, which are actually not totally universal) and peculiarities (e.g., Dutch crossed dependencies) was limited to what was interesting at that time in syntactic theory. Conversely, modern researchers are not all linguistically ignorant and are not all insensitive to the tail. Many papers (Fox, 2002, Hwa et al., 2002) have explored the extent of the tail. On a larger scale, syntax has returned to machine translation (see for example, (Koehn, 2010) Chapter 11, Venugopal et al. (2007), and Ambati et al. (2009)); and attempts are being made to include semantics as well (Baker et al., 2009).

3.2 Construction Grammar: form and function in language

Construction Grammar (Goldberg, 2006, Croft, 2001, Kay, 1997) has amazingly wide appeal to linguists in a variety of theoretical frameworks from cognitive and functional linguistics to generative grammar. Constructions are pairs of form and meaning (function). They can be as simple as a word and a word sense, but the form and meaning can also be complex. For example, Kay and Fillmore (1999) describe an English construction `What is X doing Y?` that expresses incongruity (which is relevant to much recent work on sentiment detection) and includes sentences like *What is this fly doing in my soup?* and *What is he doing going to the movies when he has homework to do?*.

Some constructions are part of core grammar. `S` → `NP VP` is a construction that expresses that a VP is predicated of an NP. Other constructions are peripheral. For example `Why not Verb?` (e.g., *Why not go?*), an English construction for making suggestions, does not have a Subject or a finite verb (a verb marked for tense), both of which are normally required for English sentences. Some constructions that look normal may have special meanings. *The lion is a fierce beast* has perfectly normal syntax, but it can be interpreted as a generic statement or it can be about a specific lion.

The shocking revelations of Construction Grammar are how much of language is in the periphery, and how many constructions in the core have idiosyncratic or unpredictable meanings. Constructions make up a large part of the long tail. They are like lexical items in that a few hundred of them are common, including existentials, comparatives, modals, and rates (*40 miles per gallon*) (Fillmore et al., To Appear). However, there are many more rare constructions, including things like *what a nice dress*, *if only I had gone*, *never have I seen such a mess*, *what is he doing going to the movies*, and so on. Identifying rare constructions is as important as identifying rare lexical items; the collective weight of them is too much to be ignored.

3.3 A Construction Grammar approach to modality and definiteness

This section focuses on some constructions for reasonably frequent concepts: definiteness and obligation. The sentences below express obligation in English (Example (6a)), Japanese (Examples (6b, c)) (Fujii, 2004), and Hindi (Example (6d)).[3] Although these are not sentences with exotic meanings, the constructions are very different from each

[3] Examples provided by Alok Parlikar and Vamshi Ambati.

other. The Japanese sentences sound like circumlocutions in English, but they are conventional, fixed expressions in Japanese and are quite normal. Notice that the syntax of the constructions in the three languages is not at all the same. English uses its auxiliary verb system to express modality. Hindi uses a dative case marker on the Subject and a verb whose literal meaning is to fall or befall, but indicates obligation when used with an infinitive complement and a dative case Subject. In Example (6b), the Japanese word that means *read* is in a relative clause. In Example (6c), it is in an embedded clause markerd with the complementizer *to* that modifies *ikenai* “can’t go”.

(6) a. You should read.

b. Yonda hoo ga ii yo.
read.PAST alternative NOM good REPORTATIVE
You should read.
(Literally: The alternative that you read is good.)

c. Hayaku yomanai to ikenai yo.
soon read.NEG QUOTATIVE go.COND.NEG REPORTATIVE
You must read soon.

d. Aap ko paRhnaa paRaa.
you DAT read.INFIN fall.PERF
You had to read.
(Literally: Fell to you to read.)

Definiteness (e.g. *the book* vs. *a book*) is a complex semantic notion, including concepts of identifiability, referentiality, familiarity, specificity, and uniqueness (Lyons, 1999, Payne, 1997). Definiteness does not correlate perfectly with the use of the words *the* and *a* in English (unless 80 percent is your idea of perfection). To emphasize this point, Croft (1990) lists 11 referential functions of noun phrases in English and French, two Western European languages in close long-term contact with each other. English and French use the equivalent articles (*the* corresponding to *le/la/les*, *a* corresponding to *un(e)*, or null corresponding to null) for only four of the 11 functions. Croft’s examples are shown in Table 1. A description of the data in terms of Construction Grammar would consist of pairs of form and meaning. For example, for English, the referential function *generic of a count noun* is expressed by the form *plural noun with no article*, whereas in French, it is expressed by the form *plural noun with definite article*.

Languages that do not have definite and indefinite articles show even more divergences from English. The functionalist school of linguistics (Comrie, 1981) points out that in the most prototypical sentence the agent is definite and the patient is indefinite. When the oppo-

Reference Type	**English**	**French**	
Specific indefinite	He broke **a vase**	Il a cassé **un vase**	same
Specific definite	He broke **the vase**	Il a cassé **le vase**	same
Proper name	The concert will be on **Saturday**	Le concert sera **samedi**	same
Specific manifestation of an institution	He went to **the bank**	Il est allé à **la banque**	same
Partitive of a mass noun	I drank **wine**	J'ai bu **du (=de le) vin**	different
Generic mass noun	The French love **glory**	Les Français aiment **la gloire**	different
Specific manifestation of an abstract quality	He showed **extreme care**	Il montra **un soin** extrême	different
Generic of a count noun	I love **artichokes**	J'aime **les artichauts**	different
Generic of a count noun, indefinite number	Birds have **wings**	Les oiseaux ont **des (= de les) ailes**	different
Predicate nominal	He became **a soldier**	Il est devenu **soldat**	different
Specific but indefinite number of a count noun	**Dogs** were playing	**Des (= de les) chiens** jouaient	different

TABLE 1 **Comparison of form and function for articles in French and English (Croft, 1990:6-7)**

site occurs (indefinite agent or definite patient), various constructions are used to emphasize that the situation is not prototypical. The constructions may include a difference in case marking of definite Objects (Turkish specific indefinite Objects, Hebrew definite Objects), a difference in agreement with definite Objects (Mapudungun, Hungarian, Swahili), the use of an existential construction for indefinite Subjects (Chinese), or a deviation from canonical word order (Hindi, illustrated below; Chinese).

The examples below show the effect of word order change on the interpretation of definiteness in Hindi. In Example (7a) the Subject and Object are in canonical positions (Subject-Object-Verb) and can be interpreted as definite or indefinite.[4] In Example (7b), the Object and recipient come before the Subject, and the Object (*necklace*) can only be interpreted as definite. In Example (8a), a Subject (*lion*) that is not in initial position is interpreted as indefinite. In fact, the sentence has the feel of an existential sentence. In Example (8b), the Subject in initial position is interpreted as definite.[5]

(7) a. Sunaar-ne anu-ko haar bhejaa.
goldsmith-ERG Anu-DAT necklace-NOM send-PERF
The/a goldsmith has sent Anu a/the necklace.

b. Anu-ko haar sunaar-ne bhejaa.
Anu-DAT necklace goldsmith-ERG send-PERF
The goldsmith sent Anu the necklace.

(8) a. Jangal mein sher hai.
forest in lion be-PRES
There is a lion in the forest.

b. Sher jangal mein hai.
lion forest in be-PRES
The lion is in the forest.

This section closes with some remarks about machine translation and the possible uses of Construction Grammar. Baker et al. (2009, 2010a,b) propose a framework for Semantically Informed Machine Translation (SIMT), which uses semantically annotated syntactic trees in a syntax-based statistical MT system. Two semantic issues were addressed, named entities and modality. Modality was recognized constructionally

[4] From Mohanan (1994) with adjusted romanization by Dipti Sharma.

[5] Examples (8a, b) are from Mahesh et al. (2005), numbers 6 and 7.

in English with a rule-based modality tagger implemented in TSurgeon (Levy and Andrew, 2006). TSurgeon rules were used to identify modal constructions in English and annotate the syntactic tree as shown in Figure 2. The semantically annotated syntactic trees were then used to train a syntax-based MT system. The preliminary experiment described by Baker et al. (2010a) resulted in a small increase in BLEU score from 26.4 to 26.7. We cannot conclude much from this small increase. However, the approach seems promising for two reasons. First, it shows that semantic information can be useful in MT without resorting to a complex interlingua representation. Second, it was only necessary to annotate the English side of the training data with semantic information, which makes the approach viable for translation between English and low-resource languages.

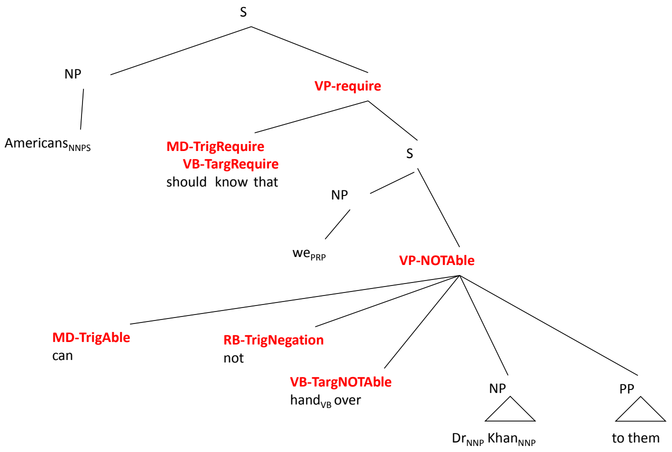

FIGURE 2 **Semantically Annotated Phrase Structure Tree**

The SIMT approach may be useful for other semantic notions such as definiteness, which have vexed machine translation researchers. Chinese, like Hindi, does not have words corresponding to *the* and *a*, although it has demonstratives (words corresponding to English *that*) and can use the number one to emphasize indefiniteness. A recent unpublished study (Flanigan et al., 2010) estimates that the choice of *the*, *a/an* or *other/none* is about 75 percent correct in a mature Chinese-to-

English MT sytem. If such a ceiling exists in a form-based SMT sytem, it may be the case that more comprehensive analysis in terms of form and function should be pursued. Chen (2004), for example, presents a linguistic comparison of the semantic functions of English determiners and various constructions in Chinese such as existential constructions with *yoo* ("have"), preverbal direct objects with the particle or co-verb *ba*, topicalization, and prenominal use of *yi* ("one"). Such an analysis could be incorporated into semantic tree annotations of the type used for modality in the SIMT approach in order to better model how the Chinese and English constructions correspond.

The phenomena discussed in this section are relevant to the study of divergences in machine translation — sentences for which simple reordering of sister nodes does not result in a good translation. The vastly different expressions of modality and definiteness in English, Japanese, and Hindi defy any small inventory of divergence types (Dorr, 1994, Mel'čuk and Wanner, 2006). The examples presented here indicate a preference for linguistic constructional approaches (Fillmore et al., To Appear) or machine learning techniques that allow for asynchronous or non-isomorphic trees between languages (DeNeefe and Knight, 2009, Gimpel and Smith, 2009, Harbusch and Poller, 2000).

4 Error Analysis

4.1 Fear of data

Proponents of statistical approaches love large data sets, but most seem to be afraid to touch the data with their bare hands, prefering instead to handle it with models and automatic scoring metrics. Many people have worked with large Chinese-English and Arabic-English data sets for decades without learning the basic syntax of Chinese or Arabic. Error analysis by humans is not unheard of (Kulick et al., 2006), but is less common than it should be. The people who say they love data the most seem to be the most afraid of looking at it.

What is the source of data avoidance? Chinese and Arabic writing systems can be an obstacle, but they can be overcome by transliteration or a bit of studying. Perhaps the issue is strong faith in statistical methods, which is sometimes justified. However, it is more likely that researchers see languages as black boxes because they lack meta-linguistic knowledge about how languages are structured.

The previous sections of this paper suggest some ways to make sense of the form and function of simple sentences consisting of Subjects, Objects, and verbs. This section proposes a fun approach to more complex sentences based on old-fashioned transformational grammar. The point of this is that complex sentences are not random distortions of simple sentences. They are the product of the interaction of well-defined simple phenomena. Many of the simple phenomena are idiosyncratic constructions, but many are regular processes that fall into a small number of sub-systems including transitivity alternations (changes in a verb's arguments), complementation (embedded clauses that are arguments of verbs), filler-gap constructions, coordinate structures, and ellipsis. After you learn some basic constructions, complex sentences become puzzles or *language knots* that are fun to untie, and reveal the inner workings of syntactic structure. If you don't come to enjoy disentangling complex sentences, you will at least develop respect for treebankers, who are able to disentangle them.

4.2 Language knots

By knowing many constructions, linguists can untie *language knots* in sentences that contain several interacting constructions. The term *language knot* was introduced in a post by David Beaver on Language Log.[6] Beaver designated Example (9a) as a *stripped cleft sluice*. Stripping, Clefting, and Sluicing are all names of constructions. The names suggest processes, referring back to the days of old Transformational Grammar (1957 to 1978 or so) when sentences were derived from more basic sentences known as deep structures by sequences of meaning-preserving tree-to-tree mappings known as transformations. The transformations often had colorful names (like Stripping, Clefting, and Sluicing) and syntax students were trained to discover the derivation of knotted sentences from their deep structures. In Example (9), *but where* (Example (9a)) is derived from the deep structure *but where they say it is changing* (Example (9b)) via Stripping, Clefting, and Sluicing.

(9) a. This time it is no longer what brands say that is changing, or how they say it, but where.

b. . . . but where they say it is changing.
Deep Structure

c. . . . but it is where they say it that is changing.
Cleft: it is . . . that . . .

d. . . . but where they say it.
Stripped: Delete all but one constituent. In this case the only remaining constituent is the headless relative clause "where they say it."

e. . . . but where.
Sluiced: Delete the remainder of a clause after a wh-word.

[6] `http://itre.cis.upenn.edu/~myl/languagelog/archives/004125.html`

During the 1970s long lists of specialized transformations gave way to theories about what kinds of transformations are possible in human language. At that point, although linguistic theories became more interesting, a larger time investment was required to understand them. Another innovation of the 1970s and 1980s was the non-derivational computationally oriented theories such as Lexical Functional Grammar, Head Driven Phrase Structure Grammar, Combinatory Categorial Grammar, and Tree Adjoining Grammar. In these frameworks, sentences are not derived from each other, but are each built independently by different lexical and grammatical choices. However, in spite of the fact that the old transformations are obsolete, many constructions are still known by their old transformation names, and modern linguists talking to each other informally still use transformations as a metaphor for linguistic constructions. Transformations may still provide simple, understandable framework for understanding the structure of complex sentences.

Example (10a) (from the Brown Corpus portion of the Penn Treebank) could be described as a topicalized relative clause with across-the-board extraction inside a passive clause with *it* -extraposition, along with a coreference anomaly (between *he* and *man*) that is caused by the change in word order (Mohanan, 1984).

(10) a. It has been truly said that anything man can imagine he can produce or create.

b. NP truly said that he can produce or create anything man can imagine.

- This is roughly the deep structure. Note that *he* can't refer to "man".
- *Anything man can imagine* is a noun phrase containing a relative clause. The derivation of the relative clause is not shown here.

c. NP truly said that anything man can imagine he can produce or create.

- *Anything man can imagine* has been topicalized inside the lower clause.[7]
- *He* can now refer to *man*.

d. That anything man can imagine he can produce or create has been truly said.

[7] It is also possible that the deep structure is more complex: *NP truly said that he can produce anything man can imagine or create anything man can imagine*. In this case, the topicalization of *anything man can imagine* involves *across-the-board* extraction from two conjoined verb phrases.

- The main clause is now passive with a sentential subject, *That anything man can imagine he can produce or create.*

e. It has been truly said that anything man can imagine he can produce or create.

- The sentential subject of the main clause is *it*-extraposed.

The next example illustrates the interaction of transitivity alternations (specifically passivization), filler-gap constructions (specifically relative clauses), and complementation under the verbs *expect* and *fear*.

Below are two sentences from the English side of the NIST MT 2009 corpus (Simpson et al., 2008). Each one has a passive clause within a passive clause within a filler-gap construction. They were found in the first 1 percent of the corpus, about 2000 words, indicating that this level of complexity is not rare (and justifying many homework assignments in introductory syntax classes). The structures shown here are produced by the parser described by Miller et al. (2000), with flattening of NP and VP constituents.

(11) Field hospitals have been set up; the UAE government will build a modern hospital which is expected to be completed in five months.

(TOP (S (NP (NNP Field) (NNS hospitals)) (VBP have) (VBN been) (VBN set) (PRT (RP up)) (: ;) (S (NP (DT the) (NNP UAE) (NN government)) (MD will) (VB build) (NP (DT a) (JJ modern) (NN hospital) (SBAR (WDT which) (S (VBZ is) (VBN expected) (S (TO to) (VB be) (VBN completed) (PP (IN in) (CD five) (NNS months))))))) (. .)))

(12) On the road between Jordan and Baghdad, two Moroccan diplomats disappeared who are feared to be kidnapped.

(TOP (S (PP (IN On) (DT the) (NN road) (PP (IN between) (NNP Jordan) (CC and) (NNP Baghdad))) (PRN (, ,) (S (NP (CD two) (JJ Moroccan) (NNS diplomats)) (VBD disappeared) (SBAR (WP who) (S (VBP are) (VBN feared) (S (TO to) (VB be) (VBN kidnapped)))) (. .)))

The transformation-style derivation of Example (10) is shown in Example (13). The derivation of Example (12) is similar.

(13) a. NP complete [$_{\text{NP}}$ which]
basic clause

b. [$_{\text{NP}}$ which] be completed
passivization of "complete"

c. NP expect [$_{\text{S}}$ [$_{\text{NP}}$ which] to be completed]
complementation under "expect"

d. [$_{\text{NP}}$ which] is expected [$_{\text{S}}$ to be completed]
passivization of "expect"

e. [$_{\text{S}}$ [$_{\text{NP}}$ which] [$_{\text{S}}$ [$_{\text{NP}}$ e] is expected [$_{\text{S}}$ to be completed]]]
*Relativization results in a string-vacuous adjunction of the wh-phrase to the sentence, leaving a trace ([*NP *e]) in its original position.*

Statistical MT systems do not explicitly model the systems of transitivity alternations, clause complementation, and long-distance dependencies. Failure to model transitivity alternations like the passive leads to inconsistency in translation of semantic roles (who did what to whom). Failure to model clause complementation also results in loss of information about semantic roles. Linguistic theories of complementation explicitly represent *you* as the Subject of both *intend* and *vote* in *You intend to vote for yourself*, a fact which may be relevant for translation into some languages. Steedman (2008) reveals systematic errors in the translation of filler-gap constructions. He used Google Translate to translate simple sentences from English to Arabic and back to English. Each input sentence had a gap in subject position (*the company that [*NP *e] bought the bank*) or in object position (*the company that the bank bought [*NP *e]*). Regardless of the input, Google Translate favored output with Subject gaps (*the company that bought the bank*). There is a reversal of semantic roles when *the company that the bank bought* is translated as *the company that bought the bank*. If you conduct your own experiments with Google Translate using Arabic or some other language, you may find that you sometimes get correct results even for gaps in object position. SMT is not always wrong, but it is strange that it can't be right more often when there are recognizable patterns in well-understood constructions.

To end this section on an optimistic note, the following sentences were all translated well by Google Translate at the time this article was written. (Note that there is a correctly translated object-gap construction in Example (14a), and possibly also in Example (14b) depending on the analysis of the complement of *expect*.) In each example, the first sentence was input to Google Translate and translated into Arabic. The second sentence is the result of translating back into English. The Arabic translations are not included here.

(14) a. **Input:** I saw the bridge that the customer expects the company to complete.
Output: I saw the bridge that the client expects the company to complete.

b. **Input:** I saw the bridge that the customer expects to be completed.
Output: I saw the bridge that the client expects to be completed.

c. **Input:** I saw the bridge that is expected to be completed.
Output: I saw that the bridge, which is expected to be completed.

5 Conclusion and Recommendation

The recommendation that has been emphasized throughout this paper is for language technologists to understand the object of study, human language. The paper has focused on the variety and complexity of human languages and has also emphasized the importance of both regularity and idiosyncrasy. Variety exists in the tendency to use word order, case marking, or agreement as the primary mechanism for grammatical encoding. It is also manifest in the diversity of constructions that are used to express notions like modality and definiteness. Complexity arises from the interaction of simple sub-systems, as well as from the existence of many rare, idosyncratic constructions. It may be the case that statistical models and machine learning methods will eventually capture everything described in this paper. But in order to understand where current methods are falling short, we as a field need to understand the data.

6 Acknowledgements

I would like to thank the following people for their comments on this paper and help with foreign language examples. All mistakes are mine. Thank you, Greg Hanneman, Carol Tenny, Stephan Vogel, Jeff Flanigan, Vamshi Ambati, Alok Parlikar, Dipti Sharma, Rajesh Bhatt, Patrick Littell, Ken Church, Chris Callison-Burch, the editors, and the reviewers.

References

Ambati, Vamshi, Alon Lavie, and Jaime Carbonell. 2009. Extraction of syntactic translation models from parallel data using syntax from source and target languages. In *Proceedings of MT Summit XII*. Ottawa, Canada.

Baker, Kathy, Steven Bethard, Michael Bloodgood, Ralf Brown, Chris Callison-Burch, Glen Coppersmith, Bonnie Dorr, Nathaniel Filardo, Kendall Giles, Anni Irvine, Mike Kayser, Lori Levin, Justin Martineau, Jim Mayfield, Scott Miller, Aaron Phillips, Andrew Philpot, Christine Piatko, Lane Schwartz, and David Zajic. 2009. Semantically informed machine translation (SIMT), final report of the 2009 summer camp for applied language exploration. Tech. rep., Human Language Technology Center of Excellence, Johns Hopkins University.

Baker, Kathy, Michael Bloodgood, Bonnie Dorr, Nathaniel Filardo, Lori Levin, Scott Miller, and Christine Piatko. 2010a. Semantically-informed machine translation: A tree-grafting approach. In *Proceedings of the Association for Machine Translation in the Americas (AMTA)*. Denver.

Baker, Kathy, Michael Bloodgood, Bonnie Dorr, Nathaniel Filardo, Lori Levin, and Christine Piatko. 2010b. A modality lexicon and its use in automatic tagging. In *Proceedings of the Language Resources and Evaluation Conference*. Malta.

Bender, Emily M. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32. Athens, Greece: Association for Computational Linguistics.

Bresnan, Joan. 2000. *Lexical Functional Syntax*. Blackwell.

Bresnan, Joan, Ronald Kaplan, Stanley Peters, and Annie Zaenen. 1982. Cross-serial dependencies in dutch. *Linguistic Inquiry* 13(4):613–635.

Chen, Ping. 2004. Identifiability and definiteness in chinese. *Linguistics* pages 1129–1184.

Comrie, Bernard. 1981. *Language Universals and Linguistic Typology*. The University of Chicago Press.

Croft, William. 1990. *Typology and Universals*. Cambridge University Press.

Croft, William. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. Syntax and Semantics, volume 34. Academic Press.

DeNeefe, Steve and Kevin Knight. 2009. Synchronous tree adjoining machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 727–736. Singapore.

Dorr, Bonnie. 1994. *Machine Translation: A view from the lexicon*. MIT Press.

Fillmore, Charles, Ressell Lee-Goldman, and Russell Rhodes. To Appear. The framenet constructicon. In H. Boas and I. Sag, eds., *Sign-based Construction Grammar*. Center for the Study of Language and Information (CSLI).

Flanigan, Jeff, Vamshi Ambati, Stephan Vogel, and Lori Levin. 2010. Determiners in Ch-En MT. Unpublished PowerPoint presentation.

Fox, Heidi J. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 304–311.

Fujii, Seiko. 2004. Lexically (un)filled constructional schemes and construction types, the case of Japanese modal conditional constructions. In

M. Fried and J.-O. Östman, eds., *Construction Grammar in a Cross-Language Perspective*, pages 121–155f. John Benjamins Publishing Company.

Gimpel, Kevin and Noah A. Smith. 2009. Feature-rich translation by quasi-synchronous lattice parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 219–228. Singapore.

Goldberg, Adele E. 2006. *Constructions at Work: The nature of generalization in Language*. Oxford University Press.

Greenberg, Joseph H. 1963. *Universals of Languages*. MIT Press.

Harbusch, Karin and Petr Poller. 2000. Non-isomorphic synchronouns TAGs. In A. Abeillé and O. Rambow, eds., *Tree Adjoining Grammars: Formalisms, Linguistic Analysis and Processing*. CSLI.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie. 2005. *World Atlas of Language Structures*. Oxford University Press.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, and O. Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the ACL*.

Iggesen, Oliver A. 2005. Number of cases. In M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie, eds., *World Atlas of Language Structures*, chap. 49–50, pages 202–209. Oxford University Press.

Kay, Paul. 1997. *Words and the Grammar of Context*. CSLI Publications.

Kay, Paul and Charles Fillmore. 1999. Grammatical constructions and linguistic generalizations: The "what's X doing Y" construction. *Language* 75:1–33.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press.

Kulick, Seth, Ryan Gabbard, and Mitchell Marcus. 2006. Parsing the Arabic treebank: Analysis and improvements. In J. Hajič and J. Nivre, eds., *Proceedings of Treebanks and Linguistic Theories*. Institute of Formal and Applied Linguistics, Prague, Czech Republic.

Kurimo, Mikko, Sami Virpioja, and Ville Turunen. 2009. `http://www.cis.hut.fi/morphochallenge2009/`.

Levy, Roger and Galen Andrew. 2006. TRegex and TSurgeon: Tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Lyons, Christopher. 1999. *Definiteness*. Cambridge University Press.

MacWhinney, Brian. 2004. A unified model of language acquisition. In J. Kroll and A. D. Groot, eds., *Handbook of Bilingualism: Psycholinguistic approaches*. Oxford University Press.

Mahesh, R., K. Sinha, and A. Thakur. 2005. Translation divergence in English-Hindi MT. In *In Proceedings of the European Association for Machine Translation (EMAT)*.

Mel'čuk, I. and L. Wanner. 2006. Syntactic mismatches in machine translation. *Machine Translation* pages 81–138.

Miller, Scott, Heidi J. Fox, Lance A. Ramshaw, and Ralph M. Weischedel. 2000. A novel use of statistical parsing to extract information from text. In *In Proceedings of Applied Natural Language Processing and the North American Association for Computational Linguistics*.

Mohanan, K.P. 1984. Lexical and configurational structures. *The Linguistic Review* 3(2).

Mohanan, Tara. 1994. *Argument Structure in Hindi*. Center for the Study of Language and Information.

Payne, Thomas. 1997. *Describing Morphosyntax: A guide for field linguists*. Cambridge University Press.

Sgall, Petr, Eva Hajičová, and Jarmilá Panevova. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht, Holland: Reidel.

Simpson, Heather, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, and Boyan Onyshkevych. 2008. Human language technology resources for less commonly taught languages. In *Proceedings of the LREC 2008 Workshop on Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages*.

Steedman, Mark. 2008. On becoming a discipline. *Computational Linguistics* 34(1).

Venugopal, Ashish, Andreas Zollmann, and Vogel Stephan. 2007. An efficient two-pass approach to synchronous-CFG driven statistical MT. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 500–507. Rochester, New York.

Ward, Gregory and Betty J. Birner. 2004. Information structure. In L. R. Horn and G. Ward, eds., *Handbook of Pragmatics*, pages 153–174. Oxford: Basil Blackwell.