Fiat Lingua

Title: The Linear Aspects of Syntax: Ideas for Your Conlangs

Author: Douglas Ball

MS Date: 07-23-2012

FL Date: 10-01-2012

FL Number: FL-00000D-00

Citation: Ball, Douglas. 2012. "The Linear Aspects of Syntax: Ideas for Your Conlangs." FL-00000D-00, *Fiat Lingua*, <http://fiatlingua.org>. Web. 01 Oct. 2012.

Fiat Lingua is produced and maintained by the Language Creation Society (LCS). For more information about the LCS, visit http://www.conlang.org/

The Linear Aspects of Syntax: Ideas for Your Conlangs* †

Douglas Ball

Abstract

As part of an effort to encourage conlangers to explicate the syntaxes of their languages, this paper discusses several of the most common linear order generalizations found in natural languages. Among those discussed are the linear order generalizations surrounding heads, the order of verbal arguments, ordering of elements with certain information statuses, and ordering by "weight".

1 Introduction

In many descriptions of conlangs on the web, the section on word order (or syntax, more generally) is rather short and not all that detailed. While this could have a number of different causes, this paper operates on the premise that providing more information about how (a subpart) of syntax works in natural languages could help conlangers to find ways to extend and elaborate their syntax sections.

As I see it, this paper could be utilized in one of two ways. For conlangers interested in creating natural language-like conlangs, this paper provides details that are ripe for "importation" into your conlang(s). For conlangers interested in stretching the boundaries of language—whether through alien languages or interesting thought experiments—this work will provide you with a primer of what *not* to do. In addition, it gives you some language parameters that you can play with to alter and stretch the boundaries of what we know and think about language.

*This paper is a non-trivially revised version of my talk "Conlanging and the Linear Aspects of Syntax" from the first Language Creation Conference in 2006. I appreciate the questions and the comments raised by the participants of that conference that ultimately improved this paper. Originally, this talk also discussed the ideas of linearization theory, as found in frameworks like some versions of Head-driven Phrase Structure Grammar (HPSG) (Sag, Wasow, Bender 2003 is a general introduction; Gazdar and Pullum [1981] offer an accessible introduction to linearization theory), some versions of Lexical-Functional Grammar (LFG) (Bresnan 2001), and Categorial Grammar (particularly as proposed by Dowty 1995). While these theories still inform this paper, the original theoretical discussion has been removed.

†This paper was originally put together in summer 2009. In revising it in 2012, I have just corrected minor errors, but otherwise left it as is. This also means that the examples of Skerre are a bit dated, but they remain as a documentation of this particular version of the language.

1.1 Why Linearization?

The question arises why I should be focusing on the linear aspects of syntax. This focus derives from what seem to be some basic properties of all natural languages. The limits on what natural languages can package into a single word (which exist for some currently unknown reason) necessitate that the information be spread out temporally and, more often than not, across many words. Consequently, all natural languages have multiword expressions. In other words, there is no language where a word like *havem* means "We were surprised that they came over to our house last weekend for the party." Certainly, there are languages like the Caddoan language Wichita that can jam quite a bit into a word, as exemplified in (1):

(1) Wichita
kiyaːkíriwáːcʔarasarikitaʔahíːriks
kiyaːki- riwaːc- ʔaras- (r)a- ri- kita- ʔa- hiríːk- s
QUOT.AOR.3- big- meat- COLL- PORT- top- come- ITER- IPFV
'He brought the large quantity of meat up to the top.' (Rood 1976, 65)

But even in Wichita, while sentences can potentially be just one word, not all of the sentences are just one word long. To express particular meanings, more than one word must be used. This can be seen in (2), which includes the word in (1) in a sentence:

(2) Wichita
waːcʔarʔa kiyaːkíriwáːcʔarasarikitaʔahíːriks niyaːhkwírih.
waːcʔarʔa kiyaːki- riwaːc- ʔaras- (r)a- ri- kita- ʔa- hiríːk- s na-
squirrel QUOT.AOR.3- big- meat- COLL- PORT- top- come- ITER- IPFV PTCP-
yaːk- r- wi- (h)irih
wood- PL- stand.upright- LOC
'Squirrel brought the large quantity of meat up to the top of where the tree was.' (Rood 1976, 266)

So as long as there are sequences of words in natural language, there has to be some sort of linear order to them. And in fact, the discipline of syntax, as modernly conceived, is at least partly devoted to characterizing the linear orders in natural languages and trying to come up with explanations for their patterning.

And, in fact, most syntactic theories utilize more elaborated data structures than just the "pure" linear order in their proposed understanding of natural language syntax. But to keep the discussion at a more basic level, I will focus on linear order generalizations and leave a full elaboration of why the generalizations are what they are for another venue. I encourage interested readers, though, to seek out works that do discuss possible explanations of generalizations; some places to look include those given in the references section.

1.2 Plan of the Paper

This paper, however ambitious it may seem to be, cannot tackle the whole of the problem of linear ordering in natural language in one fell swoop. Therefore, I focus on four "case studies" of linearization generalizations. These generalizations cover some of the most pervasive and common linear order patterns in natural languages, and so, I hope, will be of the most interest to conlangers.

The first concerns the linearization of the central word of a group of words—what is called the "head"—and the other words related to it, its dependents. The second concerns the linearization of the dependents, in particular a subject known as arguments, internal to themselves. The third concerns the role of information structure—how new or old to the discussion a linguistic expression is—in linear ordering generalizations. Fourth, I discuss the interesting behavior of "heavy" phrases—expressions with lots of words—within clauses.

2 Heads and Linear Ordering

A notion that has emerged as a particularly useful one for the study of syntax—one that occurs in various guises in various frameworks—is the notion of "head." And as it turns out, one use for the notion of head is in linear ordering generalizations. So I begin the "case studies" by examining heads and their linear order generalizations.

2.1 About Heads

What exactly, though, is a head? As the metaphorical use of the body part term "head" indicates, the head is the central part of a group of words. This term, like so many in linguistics, has its controversies, and in some cases, analysts are divided about exactly what the correct head is (or if a single one should be recognized). There are, though, some reasonably reliable heuristics for determining what the head is (if any) among a group of words.[1]

Trask (1993, 125) defines a head as "[t]hat element of a constituent that is primarily responsible for the syntactic character of the constituent." This is probably not the most transparent definition, so let me consider what it means in more depth. If this element is primarily responsible for the syntactic character of its larger unit, it stands to reason that if a word requires other words around it to have particular forms (thus, determining a part of the syntactic character of the unit), then it is probably a head. (This criterion is called the criterion of *subcategorization*.) On this criterion, verbs and prepositions (or postpositions), which, in some languages, require other words to have particular case forms, must be heads.

Furthermore, if one part of the phrase is primarily responsible for the syntactic character of the phrase as a whole and some phrases are comprised of single words and some phrases are comprised of multiple words, then the single words must be the heads of this kind of phrase (the criterion of *distributional equivalence*). Thus, in phrases like *swam* and *washed the car*, the verb must be the head, because only a single verb (like *swam*) can appear in the same syntactic locations as the more elaborated phrase (e.g. *washed the car*). Notice that *the car*, on the other hand, cannot appear appear in the same syntactic locations as *washed* or *swam*. Thus, this criterion is almost paramount to saying that the head is the only obligatory member of the phrase.

The above criteria of *subcategorization* and *distributional equivalence* give the most uniform results. However, two more criteria have been put forth. One is the *locus of morphosyntactic marking*. Basically, this criterion says that the word that is inherently inflected (that is, it does not agree with another word) for case, tense, aspect, gender, etc. is the head. However, this criterion must be treated with caution because, in some instances in some languages, the morphological exponent

[1]This discussion here has been greatly informed by the discussion in Beavers 2003. In particular, the criterion names are all taken verbatim from this source.

of case, tense, aspect, gender, etc. appears in a location defined with the respect to the edge of the phrase, not on the head (Perhaps the best known instance of this is the English possessive *'s*, which goes at the end of an NP, not on the head noun). There is also the criterion of *semantic characterization*. To apply this criterion, one looks at the semantics of the unit. Since *washed the car* describes an event of washing and not a kind of car, the head must be *washed*. This last criterion is most likely to give a result that differs from the earlier criteria.

On the whole, these criteria generally point to the same sort of words in various expressions. Thus, the head of a sentence is a finite verb (*John **has** written a book*), the head of prepositional phrase is a preposition (***in** the night*). the head of a verbal unit is the verb (***washed** the car*), and, perhaps most controversially, the head of a nominal expression is a noun (*the **computer***).[2] And it appears that roughly equivalent words behave as the heads of roughly equivalent phrases across languages, insomuch as the same phrase types recur across languages.

2.2 Head Linear Order Generalizations

The interesting linear order generalization with regard to heads is that languages tend to consistently place them at one peripheral part of the phrasal unit. Thus, phrases can either be characterized as head-initial or head-final. Furthermore, most languages are reasonably consistent across the headed phrases within the language. The relatively strong consistency underlies several of the famous word order universals proposed by Greenberg (1963).

To see the consistency, let me consider two examples. My conlang Skerre is an example of the head-initial type, as shown in (3) (heads bolded):

(3)

Phrase Type	Skerre	Translation
Clause	**Ehosi** tsa keriyos a hantsi	'The man ate the meat'
	PFV.eat ERG man ABS meat	
Adpositional Phrase	**te** wiyet	'on the boat'
	LOC boat	
Noun Phrase	**yaak** i arik-he	'father of my friend'
	father GEN friend=1SG.POSS	

In contrast, Turkish is very consistently head-final, as shown in (4) (heads again bolded):

(4)

Phrase Type	Turkish	Translation
Clause	Ben sizi **gördüm**.	'I saw you'
	1SG.NOM 2SG.ACC see.PRET	
Adpositional Phrase	altı saatten **sonra**	'after six hours'
	six hour.ABL after	
Noun Phrase	radyonun **sahibi**	'owner of the radio'
	radio.GEN owner	

(Examples from Thomas 1967, 42, 87, 60)

While a lot of languages are more or less reasonably consistent (English, for example, is pretty consistently head-initial), some languages are "mixed." One of the more famous examples is Ger-

[2]Certain frameworks, most notably late Government and Binding Theory as well as the Minimalist Program, adopt the view that the determiner is the head of nominal expressions like ***the** computer*. Beavers 2003 offers a nice discussion of the issues and reasons to think that both the determiner and the noun exhibit some head-like properties.

man (and other Germanic languages). Here, the finite verb appears in second position in main clauses,[3] while main verb appears finally in these clauses, as in (5):

(5) German
Er [**hat** [das Buch **gekauft**]].
3SG.M.NOM have.3SG DEF.ACC book buy.PPTCP
'He bought the book.' (Graves 1990, 103)

In (5), *hat* is the finite verb. It is initial in its phrase, demarcated by the outside brackets. However, the subject is even further to the left, making the verb second in the sentence. The main verb is *gekauft* in (5). It is final in its phrase, marked by the inner brackets. So, in both cases, the head appears on an edge of its phrase, but other parts of the sentence obscure this to a degree, and the two kinds of verbal groupings differ in which edge is chosen.

While German shows that languages need not be precisely uniform in their placement of heads (and, in fact, this "mixed pattern" has been quite diachronically stable, as these patterns have existed in Germanic languages for well over 1,000 years), the fact remains that most languages are relatively consistent in their placement. The places where head-initial/head-final splits occur are not currently very well-understood, but it seems that a split between main and subordinate clauses (like in German) and occasional splits between clausal and adpositional elements (see Dryer 2008 for discussion of data and some possible causes of the splits) are among the most frequent, within this relatively rare phenomenon.

3 Ordering of Verbal Arguments

Turning away from heads, let us look at the ordering generalizations related to the non-heads, specifically the subset called "arguments"[4]. A working definition of an argument is an obligatory dependent of some head. In particular, I focus on arguments of verbs, because just about every verb in every language has at least one argument, and thus, the behavior of verbal arguments has been well-studied.

3.1 Prominence Among Arguments

An important insight about verbal arguments is that, for many syntactic and semantic purposes, they are not equal. A means of understanding their unequalities is via the notion of prominence (see Levin and Rappaport Hovav 2005, ch. 6 for further discussion of this idea). The idea then is that some arguments are more prominent—semantically-salient— than others and this allows them more 'privileges', semantically and also syntactically.

The most prominent arguments are those that instigate the action: what roughly can be considered to be agents. The least prominent arguments are those that are altered or affected by the action: what roughly could be called patients. An important thing to keep in mind about the notion

[3]German syntax adds further wrinkles in subordinate clauses and questions/imperatives. In subordinate clauses, the finite verb is final and main verbs are penultimate. In questions and imperatives, the finite verb is strictly initial; there is no initial subject. For the sake of brevity, I do not discuss these further wrinkles in depth.

[4]The linguistic term derives from the term from logic. This is due to the fact that logical formalisms were used and continue to be used for describing semantics, and linguistic arguments are arguments—in the logical sense—of logical predicates.

of prominence is that it is relative; even if a particular argument is not exactly the most prototypical agent, it can still be said to be more prominent than another argument, such as a patient.

Agent and patient do not exhaust the possible relationships an argument can have with a predicate. There are also the indirectly affected arguments in three-place predicates, like the recipient in a *giving*-event. Within the linguistics literature, there is a fair amount of controversy about what level of prominence they should be given. Recipients are sometimes treated as just being below agents in prominence; other times, they are treated as being the least prominent, below even patients. The data that considered below will just feed this controversy: they give evidence in both directions.

3.2 Linear Ordering Generalizations with Arguments

Unlike the ordering of heads, the linear ordering generalizations with arguments do not appear to be quite so particular to individual languages. Instead, there is an overall tendency, given in (6):

(6) More Prominent Arguments $\prec$ Less Prominent Arguments
"More prominent arguments appear to the left of less prominent arguments"

This left-to-right order seems to hold regardless of where the head is placed. Furthermore, the tendency in (6) plays out in different ways in different kinds of languages. In (most) languages with fixed word order (along the lines of English), the order in (6) is the only order possible. In languages with freer word order (along the lines of Russian), the order shown in (6) is the most pragmatically unmarked order. Because (6) is a tendency, there are, of course, exceptions. However, many of the exceptions are cases where information status considerations come into play (see section 4). Less frequently, other considerations (such as those noted in section 5) also create exceptions to (6).

Let me illustrate (6) with two brief case studies: one involving the orders found in the Germanic family, and the other concerning the orders found in the Taiwanese Austronesian language, Seediq.

3.2.1 Case Study 1: the Germanic Family

The Germanic family is one that has both fixed word order languages and freer word order languages, so it provides an excellent opportunity to illustrate how (6) plays out in both kinds of languages. In this subsubsection, the examples will all be of *giving*-events, so there will be three arguments to worry about: the giver or Agent NP, the thing given or Patient NP, and the entity receiving the gift or Recipient NP.[5]

Within the Germanic family, the fixed word order languages require the order in (7):

(7) Agent NP $\prec$ Recipient NP $\prec$ Patient NP
"The Agent NP precedes the Recipient NP, which, in turn, precedes the Patient NP"

One such fixed word order Germanic language is Dutch. In (8), there is an Agent NP, *Jan*; a recipient NP, *zijn vader*; and a Patient NP, *het boek*. As (8) shows, these three arguments appear in the order expected from (7); the head of the subordinate clause—bolded—is final:

[5] All examples in this subsubsection are given as subordinate clauses, to illuminate the differences that would be obscured by using main clause examples.

(8) Dutch
... dat Jan zijn vader het boek **geeft**.
... that (name) his father DET book gives
'... that Jan gives his father the book' Paul Kiparsky, class handout

Fitting with the characterization of Dutch as a fixed word order language, all other orders of *Jan*, *zijn vader*, and *het boek* are unacceptable.

Moving to another fixed word order Germanic language, Swedish, we can see that the placement of the head does not matter for (7). In contrast to Dutch, Swedish allows for (and, in fact, requires) verb-medial subordinate clauses. Yet, when we consider the translational equivalent of (8) in Swedish, we get the same order of arguments as Dutch, even though the head is in a different position (second in the clause—bolded). This is shown in (9):

(9) Swedish
... att Jan **ger** sin far boken.
... that (name) gives his father book-DET
'... that Jan gives his father the book ' Paul Kiparsky, class handout

As (9) shows, Swedish has the order that (7) leads us to expect. All other orders of *Jan*, *sin far*, and *boken* are unacceptable in Swedish.

Turning to a freer word order Germanic language, we consider German. Befitting German's status as a freer word order language, the order Agent NP $\prec$ Recipient NP $\prec$ Patient NP from (7) is not required. But this order is the least pragmatically marked order in German. This means that the order Agent NP $\prec$ Recipient NP $\prec$ Patient NP can appear in the largest number of contexts and this arrangement of arguments has the fewest "extra" meanings; examples of the "extra" meanings I mean here would include the sense of "emphasis" given to a particular argument in English when it is intonationally stressed (JAN gave his father the book) or occurs after the verb BE in constructions like *It is* JAN *that gave his father the book*.

The translational equivalent of (8) and (9) also exists in German; it is given in (10) (note that German patterns with Dutch and has the phrase-final head—bolded in (10)—in subordinate clauses):

(10) German
... dass Jan seinem Vater das Buch **gibt**.
... that (name) his.DAT father DET.NEUT.ACC book gives
'... that Jan gives his father the book' Paul Kiparsky, class handout

However, in contrast to Dutch and Swedish, all other orders of *Jan*, *seinem Vater*, and *das Buch* are possible. But only the order shown in (10) is the contextually unmarked "neutral" word order: neither *Jan*, *seinem Vater*, nor *das Buch* is being contrasted with other words and this order can appear in the largest number of contexts.

So, this brief Germanic case study seems to suggest two things about linear order in natural languages. First, the order of arguments is not particularly dependent on where the head verb is. Second, the required orders in fixed word order languages appear to mirror the pragmatically "neutral" orders of freer word order languages.

3.2.2 Case Study 2: Seediq

The Germanic data are interesting, but taken by itself, it could just be an interesting fact about that family. So, to show that some of the general principles discussed above in section 3.2.1 apply elsewhere in the world's languages, I move to a discussion of Seediq, an Austronesian language spoken in Taiwan. This is not to suggest that the syntax of Seediq is particularly close to that of the Germanic languages; quite a few of the details are significantly different. In fact, the differences here will also allow us to see an example of one general type of exception that I mentioned earlier: exceptions to (6) often are due to "conflicting" information structure considerations.

Seediq is a typical western Austronesian language. Its heads are rigidly initial in their phrases. Additionally, all of its verbs have different forms that indicate that a particular argument NP has a special syntactic status. The particular NP is called the trigger[6] and the verb forms are known as foci in the literature on Seediq.[7] In Seediq, the trigger is marked by the prenominal word *ka*. The *ka*-phrase is required to be clause-final, regardless what focus form the verb is in. This is shown in (11); the (a) example has the Agent Focus (AF) form while the (b) example has the Patient Focus (PF) form:

(11) Seediq

a. Qmita huling ka Pawan.
see.AF dog KA (name)
'Pawan sees a dog' (Holmer 1996, 58)

b. Wada=mu qtaun ka Pawan.
PRET.AUX=1SG.GEN see.PF KA (name)
'Pawan was seen by me' (Holmer 1996, 58)

Aside from the *ka* phrase, Seediq NPs follow the generalization from (6) that more prominent arguments appear to the left of less prominent arguments. Thus, in Seediq, the prominence-based order holds in just a particular region of the clause: between the verb and the trigger (and, necessarily, excludes the trigger). A further quirk (mentioned earlier) is that the recipient and patient appear in Seediq in the reverse order from the Germanic family. Therefore, instead of (7), Seediq has the order in (12):

(12) (Agent NP) $\prec$ (Patient NP) $\prec$ (Recipient NP) $\prec$ Trigger NP
"The Agent NP precedes the Patient NP, which precedes the Recipient NP, which precedes, finally, the Trigger.
The only requirement is that there must be a Trigger at the end of the sentence; all other NPs are, in principle, optional"

Seeing the full pattern in (12) is a bit tricky, because verbs in Seediq that would have the requisite four arguments—an Agent, a Patient, a Recipient, and a Trigger—are either rare or non-existent. But we can see the pattern composited from two examples. The first, given in (13), has the Agent NP (the giver, *ka Awi*) as the trigger. The other two NPs are in order Patient NP (italicized) $\prec$ Recipient NP (underlined):

[6] So called because it appears, at least, to trigger a certain verb form.

[7] Other Austronesian languages have the same phenomenon, but in the discussions of them, the term 'voice' is used instead.

(13) Seediq
Wada mege *sapah* Pawan ka Awi.
PRET.AUX give.AF house (name) KA (name)
'Awi gave Pawan a house.' (Holmer 1996, 79)

In the second (14), the Recipient NP (the gift-getter) is the trigger. The other two NPs are in the order Agent NP (italicized) $\prec$ Patient NP (both parts underlined):[8]

(14) Seediq
Bniqan=*mu* lukus mu heya
give.PRET.LF=1SG.GEN clothes 1SG.GEN 3SG.NOM
'I gave my clothes to him' (Holmer 1996, 79)

These patterns are not unique to Seediq; the same sort of ordering facts (including the placement of the trigger) are found in other rigid word order Austronesian languages, such as Malagasy (Pearson 2005).

The above brief exploration of the grammar of Seediq reveals two further things. First, the patterns found in the Germanic family do not seem to be isolated to the Germanic family; the general tendency found in (6) is also found in Seediq. Second, however, the grammar of Seediq has particular complications not found in the Germanic family: the interaction between the verb form and the trigger, and the rigidity in where the *ka*-phrase occurs.[9]

4 Information Status Ordering

So far, I have discussed the linear ordering tendencies found with respect to heads and with arguments. But another common basis of linear order generalizations in natural languages is information status. What is information status, you might wonder? It is a property of phrases, having to do with their informational position in the current conversation, speech, or written text. This boils down to whether the phrase is new information, old information, re-introduced old information, etc. The study of the information status is fraught with controversies over terms and the proper classification of the requisite information status categories. For the present purposes, I will try to avoid the bulk of the controversies (see Chafe 1976; Prince 1981; Gundel 1988; Lambrecht 1994; Engdahl and Vallduví 1996, among others, for more discussion) and deal with the simple facts surrounding only two information statuses. I will define them as in (15):

(15) **Topic**: What has been under discussion, discourse old (aka "theme")
Focus: New information, often contrastive (aka "rheme," "comment")

The linearization generalization, which appears to be universal,[10] is given in (16):

[8] Ideally, the example in (14) would have the agent and recipient as non-pronominals, because (a) pronominals independently have a tendency to appear early in clauses and (b) in Seediq, the so-called "genitive" pronouns must appear after the first item in the clause. But Holmer 1996 does not give examples that would avoid these potential confounds.

[9] In the interest of full disclosure, the *ka*-phrase ordering phenomenon in Seediq does bear some resemblance to the topic-first phenomenon found in at least some Germanic languages.

[10] There is another commonly discussed information status often called antitopic, which usually appears at the end of sentences. An example would be *He is a good linguist, that guy*, where *that guy* is the antitopic.

(16) Topic $\prec$ Focus
"Phrases with a topic information status precede those with the focus information status"

This pattern presents itself in several languages. It seems especially common in languages where verb-final order is either the most frequent or is required in declarative main clauses. I will use the Basque language of the Pyrenees to illustrate.

In Basque, it is most usual to end a clause with a main verb followed by an auxiliary. An example of this is given in (17):

(17) Basque
Hemen bizi da.
here live AUX.3SG
Verb Auxiliary
'She lives here.' (King and Olaizola Elordi 1996, 18)

Moving backwards in the clause, the focus information status element (including all content question words like the translational equivalents of *who, what, where, when*, etc.) must immediately precede the verb. An example showing both a question word and a non-question NP focus is given in (18):

(18) Basque
Nor etorri zen – *Jon* etorri zen.
who come AUX John come AUX
'Who came? John came.' (King and Olaizola Elordi 1996, 204)

In the first clause of (18), *nor* 'who' is the focus element, since the question word has to be the new element or it won't be a sensical question. So *nor* has to come before the main verb, *etorri*. In the second clause of (18), *Jon* is the new information, since it is the answer to the question. Thus, *Jon* must come before the main verb in this clause.

Continuing our Benjamin Button-esque trip through the Basque clause, the last generalization is precisely the one given already in (16): the topic element must precede the focus. An example of this sort of ordering in Basque is given in (19):

(19) Basque
Ni-ri, Jonek azaldu zidan.
I-DAT Jon.ERG explain AUX
'JON explained that to me.' (Hualde and Ortiz de Urbina 2003, 460)

In (19), *niri* 'to me' appears in an initial, intonationally-separate phrase, a common place for topics. The sense of (19) is probably more like "As for me, Jon explained that to me;" that is, returning to something of relevance to the speaker, it was Jon that explained that particular thing to that speaker.

Patterns very similar to the ones found in Basque are also found in the rigidly verb-final Turkish (Hoffman 1998), the generally verb-medial Hungarian (Kiss 1995), and the reasonably free word order Australian Aboriginal language Warlpiri (Legate 2002).

The generalization in (16) also manifests itself in verb-initial languages. Tzotzil, a Mayan language spoken in Mexico exemplifies this. In Tzotzil clauses, as in many Mayan languages, the arguments usually follow the initial verb. However, in some specialized constructions, Tzotzil allows both topic elements and focus elements to appear before the verb. When both a topic and

a focus appear before the verb, they appear in that order, in line with (16). An example of this is given in (20). Note that all Tzotzil initial topics require the pre-phrase word *a*. The context (in English translation) is given in (20a) so the relative oldness and newness of the topic and focus of (20b) can clearly be seen:

(20) Tzotzil

a. Context:
Once there was an orphan. The orphan suffered greatly. Whatever the master's children ate, they ate first. They drank first.

b. A ti prove tzeb-e sovra ch'ak'bat.
TOP DET poor girl-ENC leftovers was.given
'It was leftovers that the poor girl was given' (Aissen 1992, 51)

As the context shows, the orphan is introduced early in this (short) discourse, so the coreferential phrase *ti prove tzeb* 'the poor girl' clearly qualifies as a topic. *Sovra* 'leftovers' is clearly new to the discourse in (20b), so it qualifies as a focus. Given these information statuses, the order in (20b) is precisely expected from (16).

Thus, the Tzotzil data reveals that the generalization in (16) seems to hold when topics and foci are clearly not in their canonical positions (as you may recall, normally, all arguments follow the verb in Tzotzil), in addition to when the topics and foci seem like they are in more "normal" positions, as in the Basque examples above.

The Tzotzil data also shows the information status-based ordering can be found in all kinds of languages. Furthermore, contrasting the patterns in Tzotzil with Basque brings out an important point about the nature of information status-based ordering. In a language such as Basque with fairly rich morphology (especially either case or agreement morphology), the topic $\prec$ focus order appears "unadorned", that is without any additional specialized collection of words. When special information statuses need to be highlighted in more analytical languages, there is an additional collection of words. Besides Tzotzil, English offers an example, too, in the form of cleft sentences. In a cleft sentence like *It is linear ordering that we're talking about*, the "extra" *it is* is needed to contrast the newer information *linear ordering* with other possible bits of information.

5 Ordering by "Weight"

The final linear ordering phenomenon I want to address in this paper is what I will call ordering by "weight". The notion of "weight" deals with the amount of linguistic material in a given phrase. The key unit of "weight" is "heavy": heavy phrases either consist of a large number of words or involve a complex structure, such as several relative clauses or other subordinate clauses. For a great many of the cases, it is difficult to decide if the large number of words or the structural complexity is the defining quality of these phrases, because the two are highly correlated. (Deciding on this matter will not end up being relevant for our purposes.) In addition to heavy phrases, some linguists have also explored generalizations about light phrases, which are usually defined as single word phrases, perhaps phonologically dependent on another phrase. However, it is not so clear what is going on with light phrases and, aside from noting that they often occur near heads, I will not discuss them further in this paper.

The most well-established ordering by weight generalization is that heavy phrases appear at the end of clauses; that is $X \prec$ Heavy, where X is any non-heavy phrase.[11] This occurs in English. As (21) shows, a complex NP like *some friends that John had brought to the party* is much better at the end of the sentence, as in the (a) example, as opposed to in the usual object "slot" between the verb and the prepositional phrase (shown in the (b) example):

(21) English

a. I introduced to Mary some friends that John had brought to the party.

b. ?I introduced some friends that John had brought to the party to Mary.
(Hawkins 1990, 228)

However, this linear order generalization is not specific to English or the Germanic family. It has instantiations around the world, in languages with otherwise very different word order patterns. One such language is the Boumaa dialect of Fijian. In this language, the canonical order has the verb first and the doer of the action at the end; thus, Boumaa Fijian has Verb–Object–Subject (VOS) word order. This is shown in (22):

(22) Boumaa Fijian
E rai-ca [a gone] [a qase].
3SG see-TR ART child ART old.person
'The old person saw the child.' (Dixon 1988, 243)

Yet if the verb takes a complex argument, such as a complement clause, that argument must appear at the end of the clause. This is exactly what happens in (23), where the complement clause *ni o ira sana mai 'abati Boumaa* appears at the end of a clause:

(23) Boumaa Fijian
V-S-Heavy
E tu'u-na mai [o Tui Waini'eli] [ni o ira sa-na mai 'aba-ti Boumaa].
3SG tell-TR here ART (title) COMP ART 3PL INCP-FUT come invade-TR (place)
'Tui Waini'eli said that they would come and invade Boumaa.' (Dixon 1988, 243)

This phenomenon also occurs in Basque, providing an interesting twist to its word order. As mentioned in section 4, the main verb and auxiliary usually end a Basque clause (I call this the verbal complex in the examples below). The other arguments follow the Germanic order of (7), as shown in (24):

(24) Basque
Order: Agent–Recipient–Patient–Verbal Complex
[Ene aitak] [amari] [gona gorria] [ekarri dio].
my father.ERG mother.DAT skirt red.DET bring AUX
'My father brought mother a red skirt.' (Hualde and Ortiz de Urbina 2003, 448)

Yet when a Basque verb takes a complement clause, the complement clause appears at the end of the sentence. This is shown in (25), with the complement clause *Mikelek erlojua galdu duela*:

[11] This phenomenon sometimes is called Heavy-NP Shift, a term that dates from early transformational grammar, where it was proposed that these phrases underwent a transformation that moved them to the end of the clause.

(25) Basque
Order: Agent–Verbal Complex–Heavy Phrase
[Jonek] [esan du] [Mikelek erlojua galdu duela].
(name).ERG say AUX (name).ERG watch lose AUX.COMP
'Jon said that Mikel lost the watch.' (Hualde and Ortiz de Urbina 2003, 452)

Basque, then, is an interesting case. In English and Boumaa Fijian, when a heavy phrase appears at the end of a clause, it is still appearing where other arguments appear, even if it is not the ordinary spot for that kind of argument. However, in Basque, the heavy phrase appears at the end of the clause, which is not where other arguments appear.

The apparent switch from verb-final order to verb-medial order with a heavy phrase raises the question of whether the Basque pattern is always how verb-final languages work and whether the opposite of $X \prec$ Heavy ever occurs. In Japanese, some ostensibly heavy phrases are claimed to appear clause-initially, as in (26):

(26) Japanese

a. [Kinoo John-ga kekkonsita to] [Mary-ga] [itta].
yesterday (name)-NOM married that (name)-NOM said
'Mary said that John got married yesterday.'

b. ?[Mary-ga] [kinoo John-ga kekkonsita to] [itta].
(name)-NOM yesterday (name)-NOM married that said
'Mary said that John got married yesterday.' (Hawkins 1990, 231)

But there is some controversy over whether this is the same sort of "heavy shift" that occurs in English, Boumaa Fijian, and Basque or an independent information status-related phenomenon. Depending on what is ultimately decided about these Japanese examples (see Yamashita and Chang 2001 for an interesting discussion), the generalization about having heavy phrases at the end might not be a strict universal. Nevertheless, it appears that "heavy at the end" is an extremely common phenomenon.

6 Concluding Thoughts

In this paper, I have discussed four different areas of notable linear ordering generalizations in natural languages: ordering of heads, ordering of arguments, ordering of information statuses, and ordering by weight. Not all of these proved to have universal patterns, but I hope to have illustrated that the existing variation in natural languages appears to be relatively small. Heads seem to always be at the edge of their phrase. They most often are at the same edge throughout a given language, although some languages allow different head locations in different kinds of phrases. The arguments, on the other hand, typically are ordered in the same way—in order of decreasing prominence—regardless of the location of the head. And, interestingly, the most pragmatically unmarked orders in freer word order languages and the required orders in rigid word order languages appear to be the same (except for the issue regarding the order of recipients and patients). The information status order also exhibits a kind of prominence asymmetry, with the older, perhaps more accessible, information appearing before the newer information. Finally, heavy multiword phrases have a preponderance for appearing at the end of clauses.

This relatively narrow window of variation in natural languages, therefore, leaves a lot open for some interesting (if not completely naturalistic) constructed languages. A language that always puts its heads in the second position of the clause? A language where focus must precede topic? A language where all the light elements go to the edges, leaving the heavy stuff in the middle? All these things—and more—await possible exploration!

Yet, even if your conlanging style keeps you pretty close to natural languages, the topics covered here still leave many things to decide: where will your language(s) put its heads, will your language(s) have special topic and focus constructions, which order of patients and recipients will your language(s) employ? And as the many footnotes and hedges indicate, this paper also has just provided an introduction to the linear order phenomena in natural languages: there's more to uncover and once the facts have been sorted out, more details to decide about and to implement. The linguistics literature is one place to look for more, but I would not count out learning much about languages from primary (and near primary) sources: grammars, teaching materials, and, of course, speakers themselves. So, I hope this work has done its part in intriguing you about linear order generalizations and provides motivation for increasing the description of your conlang's (or conlangs') syntax.

References

Aissen, Judith. 1992. "Topic and Focus in Mayan." *Language* 68:43–80.

Beavers, John. 2003. "More Heads and Less Categories: A New Look at Noun Phrase Structure." In *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar*, edited by Stefan Müller, 47–67. Stanford, Calif.: CSLI Publications.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell Publishers.

Chafe, Wallace L. 1976. "Givenness, contrastiveness, subject, topic, and point of view." In *Subject and Topic*, edited by Charles N. Li, 25–55. New York: Academic Press.

Dixon, R. M. W. 1988. *A Grammar of Boumaa Fijian*. Chicago: University of Chicago Press.

Dowty, David. 1995. "Toward a Minimalist Theory of Syntactic Structure." In *Discontinuous Constituency*, edited by Harry Bunt and Arthur van Horck, 11–62. Berlin and New York: Mouton de Gruyter. (based on 1990 Tilburg Conference on Syntactic Discontinuity).

Dryer, Matthew S. 2008. "Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase." Chapter 95 of *The World Atlas of Language Structures Online*, edited by Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. Munich: Max Planck Digital Library. Available online at `http://wals.info/feature/95`. Accessed on 2009-06-17.

Engdahl, Elisabet, and Enric Vallduví. 1996. "Information Packaging in HPSG." Edited by Claire Grover and Enric Vallduví, *Edinburgh Working Papers in Cognitive Science, Volume 12: Studies in HPSG*. University of Edinburgh, 1–32.

Gazdar, Gerald, and Geoffrey K. Pullum. 1981. "Subcategorization, Constituent Order, and the Notion "Head"." In *The Scope of Lexical Rules*, edited by M. Moortgat, H. van der Hulst, and T. Hoekstra, 107–123. Dordrecht: Foris.

Graves, Paul G. 1990. *German Grammar*. Hauppage, NY: Barron's Educational Series.

Greenberg, Joseph H. 1963. "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements." In *Universals of Human Language*, edited by Joseph H. Greenberg, 73–113. Cambridge, MA: MIT Press.

Gundel, Jeanette K. 1988. "Universals of topic-comment structure." In *Studies in Syntactic Typology*, edited by Michael Hammond, Edith A. Moravcsik, and Jessica Wirth, 209–239. Amsterdam: John Benjamins.

Hawkins, John A. 1990. "A Parsing Theory of Word Order Universals." *Linguistic Inquiry* 21:223–261.

Hoffman, Beryl. 1998. "Word Order, Information Structure, and Centering in Turkish." In *Centering Theory in Discourse*, edited by Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, 251–272. Oxford: Clarendon Press.

Holmer, Arthur J. 1996. *A Parametric Grammar of Seediq*. Lund, Sweden: Lund University Press.

Hualde, Jose Ignacio, and Jon Ortiz de Urbina, eds. 2003. *A Grammar of Basque*. Berlin: Mouton de Gruyter.

King, Alan R., and Begotxu Olaizola Elordi. 1996. *Colloquial Basque: A Complete Language Course*. London and New York: Routledge.

Kiss, Katalin É. 1995. "Discourse configurational languages: introduction." In *Discourse Configurational Languages*, edited by Katalin É. Kiss, 3–27. Oxford: Oxford University Press.

Lambrecht, Knud. 1994. *Information Structure and Sentence Form: Topic, Focus and the Mental Representation of Discourse Referents*. Cambridge, UK: Cambridge University Press.

Legate, Julie. 2002. "Microparametric Non-Configurationality: The Case of Warlpiri." Invited talk, Department of Linguistics Colloquia Series, University of Wisconsin, Madison.

Levin, Beth, and Malka Rappaport Hovav. 2005. *Argument Realization*. Cambridge, UK: Cambridge University Press.

Pearson, Matthew. 2005. "The Malagasy Subject/Topic as an A′-Element." *Natural Language and Linguistic Theory* 23:381–457.

Prince, Ellen F. 1981. "Towards a taxonomy of given-new information." In *Radical Pragmatics*, edited by Peter Cole, 223–256. New York: Academic Press.

Rood, David. 1976. *Wichita Grammar*. New York: Garland Publishin.

Sag, Ivan A., Thomas Wasow, and Emily M. Bender. 2003. *Syntactic Theory: A Formal Introduction*. 2nd Edition. Stanford, Calif.: CSLI Publications.

Thomas, Lewis V. 1967. *Elementary Turkish*. New York: Dover Publications. Revised and Edited by Norman Itzkowitz.

Trask, R.L. 1993. *A Dictionary of Grammatical Terms in Linguistics*. London and New York: Routledge.

Yamashita, Hiroko, and Franklin Chang. 2001. "'Long Before Short' Preference in the Production of a Head-Final Language." *Cognition* 81:B45–B55.