Flat syntax: a simple syntactic annotation—and its theoretical implications

William Croft
University of New Mexico

1. Introduction

The ideas to be presented below began as an attempt to solve a practical problem in annotating sentences for descriptive and computational purposes. Practical tools such as sentence annotation schemes necessarily trade off theoretical complexity and/or empirical exceptions for simplicity and ease of use. This was the intention for developing the scheme presented here. In developing this syntactic annotation scheme, though, I concluded that the "flat" syntactic structure of this practical scheme might actually be the only syntactic structure that is theoretically necessary or desirable. This is the story to be told here.

1.1. Morphological annotation in descriptive linguistics

In descriptive linguistics, it is standard practice to provide a morphological annotation of a text corpus of the language being described (the object or source language). At a minimum, a morphological segmentation of the object language sentence or intonation unit[1] consists of: an interlinear gloss with translation of each morpheme into a metalanguage (here, English), either a metalanguage word or a set of abbreviations for technical labels of morpheme categories (e.g. SG for 'singular', PRS for 'present'); and a translation of the entire sentence/utterance into the metalanguage.[2] An increasingly accepted morphological annotation scheme is the Leipzig Glossing Rules.[3] Since we are focusing on the morphological annotation of the source language, we will present just the rules for that line only.

For concatenative morphemes, two segmentation, or morpheme boundary, markers are used: hyphen for affixes/compounds and equals sign for clitics (Rule 2). Example (1) is from Lezgian (Haspelmath 1993:207) and example (2) is from West Greenlandic (Fortescue 1984:127):

(1) Gila abur-u-n ferma hamišaluǧ güǧüna amuq'-da-č.
now they-OBL-GEN farm forever behind stay-FUT-NEG
'Now their farm will not stay behind forever.'

(2) palasi=lu niuirtur=lu
priest=and shopkeeper=and
'both the priest and the shopkeeper'

Infixes interrupt another morpheme, and hence are not concatenative. They require a different annotation, specifically an annotation with a beginning symbol and an end symbol: open and

[1] Or other unit of spoken language. Many published text collections, even of transcribed spoken language until recently, use "sentence" units, that is, do not fully specify intonational or other properties that segment the stream of spoken language, and use sentence-like punctuation. For purposes of explication, we will assume a "sentence-like" transcription and use only period/full stop.
[2] There may be additional lines for distinguishing orthographic, phonetic, phonological, and morphophonological structure of the utterance. We leave these aside here.
[3] http://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf, accessed 12 April 2025.

closed angle bracket respectively (Rule 9). Example (3) is from Tagalog (from the Leipzig Glossing Rules, no source given):

(3) b<um>ili
<ACTFOC>buy
'buy'

The beginning and end symbol indicates both the boundary of the infix, and the fact that the morpheme interrupted by the infix is itself a single morpheme, not two morphemes.

Another nonconcatenative morpheme that is represented in the morphological annotation of the Leipzig Glossing Rules is reduplication: the reduplicated part of the morpheme is separated from the base morpheme by a tilde (Rule 10). Example (4) is from Tagalog (from the Leipzig Glossing Rules, no source given):

(4) bi~bili
IPFV-buy
'is buying'

Of course, this isn't the end of morphological annotation: the mapping of morphological form to meaning is complex. However, this complexity is represented in the Leipzig Glossing Rules by the annotation on the interlinear gloss line. We are interested only in the annotation of morphological form, so we will leave that aside as well.

From the perspective of the segmentation of sentence form, there are two additional annotations of sentence segmentation that are implicit in the Leipzig Glossing Rules, but should be included in any description of its annotation scheme. Words, that is, morphologically free forms, are separated by spaces. (We ignore the very tricky issues in developing a cross-linguistically valid definition of 'word'; see inter alia Dixon and Aikhenvald 2003; Haspelmath 2011; Zingler 2020.) Finally, the sentence and/or utterance is separated from neighboring sentences/utterances in the text by period (full stop), or other punctuation used in writing (e.g. question mark, exclamation point) and in the transcription of spoken language (e.g. comma vs. period for nonfalling and falling intonation units).

In sum, there are boundary markers for segmentation of four types of concatenated morphological units—affixes (hyphen), clitics (equals sign), words (space) and sentences/intonation units (period and sometimes other markers). In addition, there are two types of boundary markers for two types of nonconcatenative units—infixes, which interrupt other units (open and close angle brackets) and reduplication (tilde).

This annotation scheme for language description does not indicate any syntactic structure, apart from the levels of the word and the sentence. It is possible to add part-of-speech labels to glossed text, e.g. in the Summer Institute of Linguistics fieldwork software; the part-of-speech labels are entered into the lexical entry for the word in the associated dictionary, and can be added automatically as a separate line in the glossed text. We will return to the issue of labels of syntactic units in §3.

1.2. Syntactic annotation in computational linguistics

There are a couple of reasons why descriptive linguists do not typically annotate syntactic structure in corpora, at least not beyond part-of-speech labeling. It is an added burden on top of the very

laborious task of transcribing audio recordings, making translations, and constructing interlinear glossed text. Syntactic structure, at least those based on formal syntactic theories, are quite complex. Worse, there is no consensus on what syntactic theory to use—in practical terms, what syntactic annotation to add to a text corpus—and whether the theory on which the syntactic annotation is based will quickly disappear and therefore the effort devoted to annotation will have been wasted.

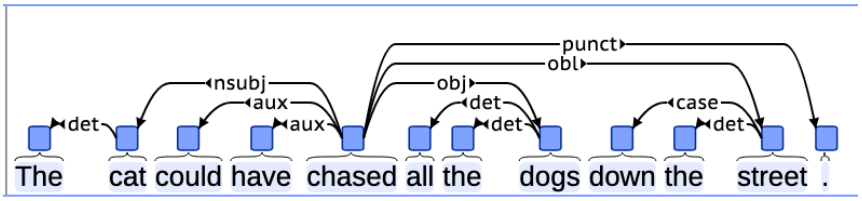
In contrast, although morphological theories have changed over time, and there are competing theories, the basically concatenative morphological model used in interlinear glossing is still widely understood and accepted, although there are problematic issues about defining words and clitics, and about the appropriate representation of noncontatenative morphology. For the latter, annotation schemes like the Leipzig Glossing Rules provide workarounds for some problematic cases for the simple morphological representation assumed by the annotation scheme.

In computational linguistics, annotation of syntactic structure is more common. A basic paradigm is to use a relatively simple syntactic annotation to annotate a corpus, and then use machine learning techniques to learn the annotation and apply it automatically to a much larger corpus; this is a type of supervised learning. Recently, there has been interest in developing a syntactic annotation scheme that is applicable across languages, not tailored to a specific language.

Computational linguists focus on practicality in annotation schemes. They are willing to exploit any additional information provided by a syntactic or semantic annotation scheme for natural language processing purposes; it does not have to conform to the latest syntactic theory, formal or otherwise. A first place to look for a practical syntactic annotation scheme, then, would be an existing, widely used cross-linguistic scheme, that is, the Universal Dependencies project.

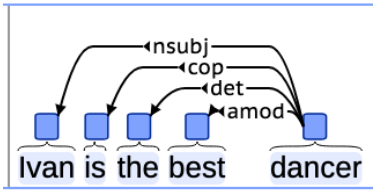
Universal Dependencies (UD; Nivre et al. 2016, 2020; de Marneffe et al. 2021), as the name implies, uses dependencies rather than the constituent structure (phrase structure) familiar to noncomputational linguists, at least in the Anglophone world, from their introductory syntax classes. Dependencies are syntactic relations between two words, one of which is the head and the other a dependent, i.e. an asymmetric or directed relation between the two words. UD labels the relations with a small set of relations developed and refined over the years. These relations are mostly familiar to linguists, though they retain a tinge of Lexical-Functional Grammar relations—LFG uses a combination of constituency and dependency to represent syntactic structure.[4] Examples (5a-c) illustrate UD dependency trees:

(5) a.

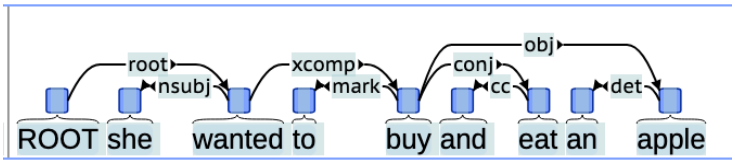


[4] https://universaldependencies.org/u/overview/syntax.html, accessed 13 April 2025.

b.



c.



In addition to this structure, UD also employs a very small set of parts-of-speech labels attached to words, and adds a larger set of morphological features similar to those used in the UniMorph project; both the parts-of-speech labels and morphological features are attached to words in the dependency structure.

UD represents a general shift in computational linguistics to dependency relations from the constituent structures familiar to many descriptive linguists thanks to American structuralism and generative grammar. The same shift to dependencies underlies a cross-linguistic semantic annotation project, Uniform Meaning Representation (UMR), which is an extended of Abstract Meaning Representation (AMR; van Gysel et al. 2021), which was developed specifically for English. UMR is partly independent of syntax, though the annotations are linked to individual words. UMR uses a graph structure, like UD, but is not restricted to dependency trees like UD is. Example (6) is an example of an AMR dependency graph:[5]

(6) The boy wants to believe the girl.

UD and AMR/UMR representations can also be represented in a text-based notation. UD uses the CoNNL-U convention, a tabular format illustrated in (7):[6]

[5] https://github.com/amrisi/amr-guidelines/blob/master/amr.md, accessed 17 April 2025.
[6] https://universaldependencies.org/format.html, accessed 17 April 2025.

(7) They buy and sell books.

```
1    They     they    PRON    PRP    Case=Nom|Number=Plur                2    nsubj
2    buy      buy     VERB    VBP    Number=Plur|Person=3|Tense=Pres     0    root
3    and      and     CCONJ   CC     _                                   4    cc
4    sell     sell    VERB    VBP    Number=Plur|Person=3|Tense=Pres     2    conj
5    books    book    NOUN    NNS    Number=Plur                         2    obj
6    .        .       PUNCT   .      _                                   2    punct
```

The rows in (7) represent each word. The columns in (7) represent the order (sequence) of words; the word; the lemma for the word; the UD part-of-speech tag; a language-specific part-of-speech tag; the UD morphological features (a list separated by |, with _ for an empty list); the word that is the parent in the dependency tree; and the UD dependency relation. (UD allows for language- or corpus-specific part-of-speech tags, and also additional parents in an enhanced dependency tree, not shown here.)

AMR's (and UMR's) graph structure is often represented in Penman notation, a text-based notation, illustrated in (8):[7]

(8) He drives carefully.

```
(d / drive-01
   :ARG0 (h / he)
   :manner (c / care-04
              :polarity -))
```

The explanation of this notation is provided below (recall that this notation is being used for semantics, hence the representation of concepts, attributes [features] and constants):

[7] https://penman.readthedocs.io/en/latest/notation.html, accessed 13 April 2025.

(9)

```
;    ┌──────────────────────── Variable (this one is the graph's top)
;    │     ┌──────────────────── Instance relation
;    ┴ ┌───┴──┐
   (d / drive-01
;     ┬ ┌──┐
;     │   └──────────────────── Concept (node label)
;     └──────────────────────── Indicates the node's concept
;           ┌──────────────────── Edge relation
;     ┌─────┴────┐
      :ARG0 (h / he)
;     ┌──┐
;        └──────────────────── Role (edge label)
      :manner (c / care-04
;                     ┌──────── Attribute relation
;                ┌────┴─────┐
                 :polarity -))
;                         ┬
;                         └──── Atom (or "constant")
```

UD (and AMR) are widely used in computational linguistics. They are rich enough representations to capture much of the syntactic (or semantic) structure that linguists as well as computer scientists are interested in. However, from the point of view of expanding syntactic annotation to descriptive linguistics, they appear to be still too complex for widespread use, especially given the resource constraints on field linguists (not to mention community members) for the vast majority of languages of the world.

The dependency structure representation, while quite powerful, is rather complex to represent. Its most readable representation is graphical, as in (5) and (6). One can use a text-based representation of a word-based dependency structure, such as the CoNNL-U format in (7) that underlies UD dependency trees or the Penman representation in (8) that underlies AMR/UMR graphs. But these text-based representations require additional "behind the scenes" notation to construct the tree or graph from a linear text-based notation. Entering annotations for a graph structure on a computer requires specialized software, and a fair amount of effort even with a good user interface. My experience working with field linguists on development of the software tool for UMR suggests that annotation with a good user interface is still quite effortful. This is partly because of linguists' unfamiliarity with the notation in either graphical or text form, and partly due to challenges in entering data structured as a dependency tree or graph. Even for text-based representations of phrase structure with square brackets, as found in the generative literature, a representation like [*A* [[*very large*] *tree*]] *fell* [*on* [*my* [*house*]]] is rather difficult to parse or even enter without errors.

Another major issue for syntactic (or semantic) annotation comes from the nature of syntactic (and semantic) representations used in both linguistics and computational linguistics: the structures are recursive. That means that the structures may be of indefinite depth—a relative clause inside an adverbial clause inside a main clause, for example—and also that a particular syntactic type, in this case a clause, must be annotated at multiple levels. The recursive nature of these

representations of syntax (and semantics) pose significant challenges for developing a simple data format that will be readable and recoverable in the distant future.

A linguistic data format that is simple and lasting is the Cross-Linguistic Data Format (CLDF) project.[8] CLDF aims to represent linguistic data in a plain text format that is valid across platforms. It uses a comma-separated value (CSV) format to create tabular data, that is, a text file where each row represents cells in a table separated by commas,[9] and a carriage return separates successive rows in the table. The type of each row and each column are also specified in the text file, using full word (not abbreviated) labels. Many rows provide metadata for the examples, or text. Here we focus only on the rows that give the analysis, gloss and translation of a text, as typically found in published descriptive data.

The CLDF format for glossed examples, also used for glossed texts, represents a typical morphological annotation of a sentence. The first row is a description of what is in each column: "Primary Text", "Analyzed Word" [morphological segmentation], "Gloss" [interlinear gloss] and "Translated Text". The following rows are lines of the text (sentences, in the case of written text), giving the sentence in the primary text; analyzed words (in the sentence); the gloss; and the free translation. The Analyzed Word and Gloss values separate words and their gosses by spaces. Users may follow the Leipzig Glossing Rules for the Analyzed Word and Gloss rows.

The CLDF format for (1), repeated as (10), is close to the text format typically displayed in publications, albeit with comma-separated values representing each version of the text line, and not formatted to align words and their glosses:

(10) a. *Typical published text format*

Gila	abur-u-n	ferma	hamišaluğ	güğüna	amuq'-da-č.
now	they-OBL-GEN	farm	forever	behind	stay-FUT-NEG

'Now their farm will not stay behind forever.'

b. *CLDF format (not including the Primary Text column)-one row, comma-separated values*

```
Gila abur-u-n ferma hamišaluğ güğüna amuq'-da-č,
now they-OBL-GEN farm forever behind stay-FUT-NEG,
Now their farm will not stay behind forever <CarriageReturn>
```

Is it possible to devise a syntactic annotation scheme that is simpler than a recursive graph structure (or phrase structure) representation, and could be easily added to a morphologically glossed text in CLDF format? And is such an annotation worthwhile for descriptive linguists as well as computational linguists? I argue here that the answer to both questions is "Yes".

[8] https://cldf.clld.org, accessed 27 May 2025.

[9] Tabs may also be used instead of commas; see https://github.com/cldf/cldf, accessed 13 April 2025. Note that tabs basically substitute for space used to separate words in interlinear glossed texts, in order to align the gloss with the source language words. Note also that UD's CoNNL-U representation is a tabular csv format.

2. Flat syntax: a very simple syntactic annotation scheme

Ideally, a simple syntactic annotation scheme would use syntactic concepts that are widely accepted; use a notation that is familiar to descriptive linguists; is relatively easy to enter on a digital keyboard; and is in a data format that is simple, transferable to new software platforms and programs, and lasting, such as the tabular csv format used in CLDF. The scheme proposed here satisfies these conditions to a great degree.

2.1. The only syntactic levels/types of units/constructions between sentence and word are clause and phrase

Interlinear glossed texts have two syntactic units or "levels", in the sense of a "higher" level construction including constructions at a "lower" level, namely sentences and words: a sentence is made up of words. Different syntactic theories propose a range of units between words and sentences. The only two that are very widely agreed on are **clauses** and **phrases**.

By 'phrases', I mean the prototypical referring phrases that are classified as subject, object and oblique, or other classifications, and usually have a head which denotes a referent, and possibly also modifiers. However, other multiword units that serve as elements of clauses are included as 'phrases'. The important point is that phrases are (sometimes multiword) units that serve as elements of clauses.

Clauses prototypically involve a predicate, a set of referring phrases (arguments in the broad sense). Clauses also contain other units, such as predicates, tense-aspect-mood-polarity "auxiliaries", and "adverbs". In *Morphosyntax* (Croft 2022), I argue that most non-argument units can be analyzed as parts of complex predicates; but that complex predicates often do not constitute a formal (syntactic) grouping of words into a contiguous phrase-like unit. Thanks to optionality, null anaphora/null instantiation, and ellipsis, not all of these elements of clauses may be present.

Sentences may also include units other than clauses, most notably a heterogeneous category of units that Heine (2023) calls 'interactives'. Heine lists the following subtypes of interactives (drawn from Heine 2023:49-50, Table 1.6):

(11) a. Attention signals: *look, hey,* etc.
b. Directives: *hang on, wait!* etc.
c. Discourse markers: *anyway, I mean,* etc.
d. Evaluatives: *good gracious, oh no!* etc.
e. Ideophones: *bang!, boing* etc.
f. Interjections: *oops, ouch* etc.
g. Response elicitors: *right? huh?* etc.
h. Response signals: *alright, certainly not* etc.
i. Social formulae: *hello, goodbye* etc.
j. Vocatives: *dad, honey* etc.

Some of Heine's interactives stand alone as utterances, while others may be parts of sentences.

In other words, sentences consist of clauses and other elements, which are often words but may be phrases (e.g. *certainly not*). Clauses consist of phrases and other elements, which are usually words, or other types of phrases (e.g. *right now*). Phrases are made up of words. For this reason, we have to broaden the description of what sentences, clauses and phrases are made up of. We will

describe sentences, clauses and phrases as **constructions**, as in construction grammar and the digital constructicon projects. These are actually construction "supertypes", in that there are many different types of sentence, clause and phrase constructions. The digital constructicons describe the parts of constructions as **construction elements**, or **CEs** (following FrameNet, which describes parts of semantic frames as Frame Elements or FEs). We will also call the parts of sentences, clauses and phrases 'construction elements'/CEs.

When we say that clauses and phrases (and sentences) are the only widely accepted syntactic units, this is not to say that there are not controversial or debatable specific cases. Many clausal constructions grammaticalize from two clauses that become gradually more closely integrated, with the predicates of the two clauses eventually forming a complex predicate and then possibly fusing together as a verb plus affix. The line between two clauses and a single clause with a complex predicate is sometimes not clear, or at best a subtle judgment. Likewise, a phrasal construction may have grammaticalized from a headless modifier phrase and a headed referring phrase in apposition that come to form a single unit. Again, the distinction between two phrase in apposition vs. a single phrase with a modifier and a head is sometimes unclear. Finally, the separation of discourse into "sentence" constructions is also a controversial choice, although in written language, groupings of clauses into sentences is usually accepted based on punctuation practices.

We make a significant simplification: we do not posit attributive phrases, made up of modifiers and admodifiers (modifiers of modifiers), such as *very tall* or *almost fifty,* as another level of syntactic unit or construction "below" (referring or "adverbial") phrases. These elements are represented as elements of the phrases they belong to, such as *very tall tree* or *almost fifty workers*. This is a practical simplification. Admodifiers are generally rare in discourse, and the creation of an additional "level" of syntactic construction adds complexity that would rarely be used in annotation. We will use another means to annotate admodification; see §3.1.

2.2. In addition to morpheme boundary markers, there are construction boundary markers (+ and #)

Interlinear glosses annotate affixes and clitics by the boundary markers hyphen and equals sign. These boundary markers separate words into morphemes. We also use boundary marker symbols to separate sentences into clauses, clauses into phrases, and phrases into words. More precisely, we use this notation to separate **sentence CEs** (often clauses), **clause CEs** (often phrases), and **phrase CEs** (words). Specifically, we use the hash symbol (#) to separate sentences into clauses and other sentence CEs, the plus symbol (+) to separate clauses into phrases and other clause CEs, and space to separate phrases into words (phrase CEs).

We are basically taking the type of notation already familiar to descriptive linguists in morphological annotation, and expanding it for syntactic annotation. One can think of it as "sentence segmentation". As noted in §1.1, space is already used to separate words in interlinear glossed texts, and period/full stop to separate sentences as a whole. The only additions we are making here is to add + to separate phrases (or more precisely, clause CEs) and # to separate clauses (or sentence CEs).[10]

An example of the syntactic boundary annotation for English, without interlinear glosses, is given in (9):

[10] In §§2.4-2.5 we will introduce two other notations for syntactic units.

(12) *The two brothers + might + lose + the game # and + exit + the competition.*

An example of the syntactic boundary annotation added to glossed text with tabs (used for space in interlinear glossing; see footnote 6), is given from Crow (Graczyk 2007:232):

(13)

éehk	huchalahúua	shoopaá-t +	kuss-chisshíi-wa-hche-k #
those	directions	four-DET	GOAL-go_back-1A-CAUS-DECL

awé	shíishiahe	shoopa-t +	kúh +	koolá-k .
season	different	four-DET	also	be_there-DECL

'I made them go back to those four directions; the four different seasons are there as well.'

In the CLDF example/glossed text format, a line or sentence is a single value, and words in the line/sentence, and the corresponding word glosses in the line are separated by spaces. In order to preserve the CLDF formatting, we combine the phrase and clause boundary markers with an analyzed word. Following English left-to-right orthography, where the sentence boundary marker period is appended to the last word of the sentence, in examples (12) and (13), and in following examples, phrase (+), clause (#) and sentence (.) boundary markers are appended to the last word of the syntactic unit. In the examples in this article, a space is put between the word and the syntactic boundary marker, but this is only for readability.

2.3. There is no recursion/embedding; there are exactly three levels above the word (phrase, clause, sentence)

Perhaps the most distinctive part of this syntactic annotation scheme is the absence of recursion or embedding. Specifically, this means that a construction of a particular type—clause or phrase, in particular—cannot have another construction of the same type as a CE. In other words, all clauses are at the same "level", and so are all phrases.

For example, coordination and adverbial subordination are annotated simply as two juxtaposed clauses:

(14) a. *after + I + ate # I + left*
b. *I + ate # and + I + left*

Also, a genitive referring phrase and its associated head referring phrase, coordinated referring phrases, and appositive referring phrases are annotated simply as two juxtaposed phrases:

(15) a. *the dog's + toy*
b. *the books + and the CDs*
c. the mayor + Tim Keller

In other words, # and + boundary markers simply distinguish one clause from the next, and one phrase from the next respectively. They do not indicate any specific relation between the neighboring clauses or phrases.

It is also the case that relative clauses, which as modifiers in phrases are usually represented as a clause embedded under a phrase, are annotated here as a clause next to a phrase:

(16) *Bilbo + found + the ring # that + Gollum + had + lost*

The important practical consequence of representing clauses and phrases as constructions but not their relations to other clause and phrase constructions in the sentence is that there are only three "levels" of representation of syntactic constructions:

(i) sentences, consisting of their CEs, which can be clauses, phrases, or words
(ii) clauses, consisting of their CEs, which can be phrases or words
(iii) phrases, consisting of their CEs, which are words

In §3.1, we will show a way to provide additional syntactic information on these three "levels". However, in §4 we will also argue that there are empirical reasons that suggest that recursion, represented by indefinite embedding of syntactic structures, may be unnecessary in representing syntactic structure.

2.4. So-called center embedding—constructions interrupted by other constructions—is represented like infixes (with { })

The annotation guideline in §2.3 only works if the clauses and phrases marked off by # and + are distinct spans of the text being annotated. This is not always the case, of course. In particular, relative clauses may interrupt the matrix clause that contains them. In some cases, phrasal modifiers may interrupt the head referring phrase that contains them. These are generally analyzed as examples of "center embedding" in phrase structure syntax. In other cases, some phrases are analyzed as "discontinuous", that is, they are interrupted by other units that do not belong to the "discontinuous" phrase.

It is likely that many if not all "discontinuous" phrases are better analyzed as distinct phrase constructions, rather than parts of a single phrase. For example, in many cases that have been analyzed functionally, the two "discontinuous" parts have different discourse functions (Croft 2022:159-64 and references therein). In these cases, it is better to analyze and annotate the two parts as two distinct phrases, as in the annotation of the Polish example in (17) (Siewierska 1984:60, cf. Croft 2001:187):

(17)

Nie!	*Piękny +*	*mają +*	*ogród.*	*Dom +*	*mają +*	*kiepski*.
no!	beautiful	have	garden	house	have	crummy

'No! They have a beautiful garden. Their house is crummy.'

In other cases, however, there is no reason to propose a split analysis and independent annotation of parts of an interrupted construction. Thus we must have a way to accommodate the interruption of one construction by another construction in our otherwise concatenative syntactic annotation scheme.

Our proposal is to annotate examples of phrases and clauses interrupting other phrases and clauses in the same way that infixes are annotated. We propose using curly braces { } to show that the enclosed phrase or clause interrupts the phrase or clause that surrounds it, just as < > are used to show that the enclosed morpheme interrupts the root or other morpheme that surrounds it.

English examples of "syntactic infixes" are given in (18)-(19):

(18) *The tree {that + fell + on {my} house} had + died + last winter.*
(19) *The {over the hill} gang + met + at the tavern.*

In (18), the relative clause interrupts the main clause, and the pronominal genitive phrase interrupts the adpositional referring phrase (cf. *on {the neighbor's} house*). In (19), the preposed somewhat idiomatic adpositional referring phrase interrupts the referring phrase serving as subject of the main clause.

Example (20), from Serbian-Croatian, is an instance of a "second position" marker is placed after the first word, rather than the first phrase (Comrie 1989:22; cf. Croft 2001:224):

(20) *taj {mi} pesnik + čita + knjigu + danas.*
that to.me poet reads book today
'The poet reads the book to me today.'

In the CLDF example/glossed text format, the boundary markers for an interrupting phrase or clause must be combined with a word entry. The opening { is combined with the first word of the interrupting phrase/clause, and the closing } is combined with the last word of the interrupting phrase/clause. This is basically the same as parentheses are annotated in English orthography.

In morphological parsing, the gloss line pulls out the infix's gloss and puts it either before or after the morpheme that the infix interrupts in the source language (see example (3) above). For practical reasons, however, we will leave the word order of the interrupting and interrupted clauses in the gloss as they are in the analyzed text line.

2.5. Multiword expressions ("words with spaces") are annotated distinctively (with ^)

Some expressions that are at least orthographically separate words function syntactically and semantically like a single word. For example, English *on top of* is often analyzed as a "complex preposition", and the combination *by and large* has a single meaning. These are called multiword expressions (MWEs) in computational linguistics. They are quite frequent, especially in languages like English and Chinese where much computational linguistic work is done.

Multiword expressions also pose issues in interlinear glossing, although there is no conventional annotation of them in the Leipzig Glossing Rules. I have seen a single word gloss align with a string of words in a source language, in order to indicate that the string of words is a multiword expression with a single meaning. However, such a solution makes it look as if the following words in the source language simply do not have a translation, rather than combining with the first word to express the meaning in the gloss.

It would be clearer to introduce an annotation in the source language annotation line that clearly indicates that a string of words in the source language corresponds to a single gloss in the gloss line. We propose using ^ to join words in a multiword expression, as in (21):

(21) *We + put + the lamp + on^top^of the filing cabinet .*

This proposal only applies to fixed, contiguous multiword expressions. There are also many semantically idiosyncratic expressions that are more flexible, such as English *break up* (*She broke it up*) and *pull strings* (*Strings were pulled to get the job for him;* cf. Nunberg, Sag and Wasow

1993). Nunberg et al. call these 'idiomatically combining expressions', and argue that these combinations are semantically analyzable; it is just that the meanings of the parts are idiosyncratically specific to just these collocations of words. Idiomatically combining expressions are annotated like ordinary syntactic combinations (*She + broke + it + up; Strings + were + pulled # to + get + the job + for him*). However, their semantic idiosyncrasy would have to somehow be represented in the gloss line.[11]

Finally, there are discontinuous combinations that together express a meaning, such as the standard written French negative construction *ne VERB pas*. This construction should be annotated in the same way as bipartite morphological elements. Leipzig Glossing Rule 8 gives two options: repeat the same gloss (in this case, NEG) for the two elements, or use a special gloss for one of the elements to indicate that it accompanies the first element.

2.6. Summary of annotation scheme

The complete annotation scheme for morphosyntactic analysis in the annotated source text line, including the morphological annotations from the Leipzig Glossing Rules, is given below:

Boundary markers - morphological

-	affix boundary
=	clitic boundary
< >	infix boundaries
~	reduplicated syllable(s) boundary

Boundary markers - syntactic

space	boundary between phrasal CEs/words, including admodifiers
+	boundary between clausal CEs
#	boundary between sentential CEs
.	sentence boundary (or one can use intonation unit boundary markers)
{ }	interrupting ("center-embedded") unit boundaries
^	fixed contiguous multi-word expression

This scheme satisfies the conditions listed at the beginning of §2. It uses syntactic concepts that are widely accepted, namely sentence, clause, phrase and multiword expression. (Of course, these are challenging to apply in particular cases, but this is not unlike word, affix and clitic.) It uses a notation that is familiar to descriptive linguists, namely boundary markers, to separate syntactic constructions of different types, including an interrupting construction notation analogous to the notation for infixes. It is relatively easy to enter on a digital keyboard, using symbols found on a standard keyboard (#, +, { } and ^). It is in a data format that is simple, transferable to new software platforms and programs, and lasting, usable in a text format data structure such as the tabular csv format used in CLDF.

[11] In the CLDF data format, multiword expressions would be entered as single words, that is, as parts joined by ^ without space; the words in the multiword expression will not be separated by commas or tabs in the csv format, and would have a single element in the Gloss line corresponding to the translation of the multiword expression.

3. Adding CE labels for additional syntactic information—maybe automatically?

In §1.1, we noted that some descriptive software allows a user to enter a part-of-speech label for words in a glossed text, although this annotation is not generally given in published glossed text collections. In §2.3, we noted that by excluding recursion and embedding, the syntactic annotation lacks information about how clauses are related to each other, how phrases are related to each other, and how relative clauses are related to the phrase they modify. We also noted in §2.1 that we did not posit a "level" for attributive phrases, in order to simplify the annotation scheme. In this section, we propose a way to annotate additional syntactic information, namely the role played by the construction element in the construction, using the three designated "levels" of constructions: sentential, clausal and phrasal.

Manual annotation of CE labels would be more labor-intensive than just adding the syntactic boundary markers + # {} ^ proposed in §2 to the analyzed text line in a corpus. CE labels belong to three levels: phrasal CEs, clausal CEs and sentential CEs. It is probably not realistic for a descriptive linguist to manually annotate CE labels for a large corpus, although software can be created to facilitate the data entry.

However, most of the information to generate the CE labels is already there in materials that a descriptive linguist would generate anyway, along with the flat syntax annotation scheme proposed in §2. In principle, one should not have to annotate information redundantly that is already present somewhere, in some form. Ideally, this information should be used to automatically generate CE labels in a three-level format. In practice, this is a highly non-trivial task, and it remains to be seen how much of CE labeling can be automated. In this section, we outline a suggested CE labeling scheme in a three-level format, and then summarize the description information (beyond the flat syntax annotation) that would allow the labeling to proceed automatically.

3.1. A suggested CE labeling scheme

Displaying labels for phrasal, clausal and sentential CEs is more complicated than just adding a single row for part-of-speech labeling as in the SIL language description software. There would be three rows for CE labeling: one for sentential CEs, one for clausal CEs, and one for phrasal CEs. Or at least, this is one relatively simple way to encode CEs at different levels. (In the CLDF format, these would be three additional columns for each row corresponding to a line of text.) The nonrecursive nature of the annotation scheme means that additional rows are never needed.

Ideally, there should be a small set of CE labels for CE roles, just as in UD there is a small set of part-of-speech tags for words. The set of CE labels introduced here is small, but not definitive.

Table 1 is the table with CE labels for a shortened version of the English sentence in (9) (we leave aside the morphological annotation in order to focus on the CE annotation):[12]

[12] I adopt the convention of putting the CE label for a multiword construction in the column for the first word of the construction. Although this means that the CE label is not necessarily above the head of a multiword construction, it seems better to have a consistent convention that applies to headed and non-headed constructions, and also applies to interrupted constructions where the CE label is on the first word of the first part (see below). I have removed the boundaries of empty cells in the Clausal CEs and Sentential CEs rows in the table for readability. In a CLDF format, there would be a symbol to indicate the empty cells for the remaining words in the construction. We suggest using underscore (_), as in UD's CoNLL-U text format.

Analyzed Text	*The*	*two*	*brothers +*	*might +*	*lose #*	*and +*	*exit +*	*the*	*competition .*
Phrasal CEs	Mod	Mod	Head	Head	Head	Head	Head	Mod	Head
Clausal CEs	Arg			CPP	Pred	Conj	Pred	Arg	
Sentential CEs	Main					Main			

Table 1. CE annotation of *The two brothers might lose and exit the competition.*

At the Phrasal CE level, the CEs of multiword constructions for **Mod**[ifier] and **Head** roles. Modifiers are not divided into subclasses (article, demonstrative, numeral, quantifier, adjective etc.). The one-word phrases are annotated "Heads" of their respective phrases; there are no modifying elements. Adding "Head" for one-word phrases may seem like an odd or unnecessary. However, in a language where single words can be headless phrases (e.g. a language where 'the red one' is simply 'red'), then it would be necessary to use Mod for such one-word phrases, to distinguish them from one-word phrases consisting of only the head. We have not treated *might lose* as a single complex predicate phrase, for reasons to be explained below.

In order to distinguish admodifiers from modifiers, we introduce an **Adm**[odifier] CE label, as in Table 2:

Analyzed Text	*a*	*very*	*large*	*tree +*
Phrasal CEs	Mod	Adm	Mod	Head

Table 2. CE annotation of *a very large tree.*

This provides a way to identify admodifiers and associate them with the modifier they modify, at least in conjunction with information about word order in the language (see §3), without adding an entire row for annotating attributive phrase structure.

Other CEs occurring at the phrasal level are independent words that are part of strategies for expressing a modification relation. These are **Adp**[osition], **Lnk** [linker], and **Clf** [classifier]. Another widely used strategy is indexation (also known as agreement). Indexes are almost always bound morphemes, but in cases where indexes are at least orthographically separate words, **Idx** [index] can be used to label them.

Overt conjunctions represent another strategy, used to link together modifiers, phrases and clauses; hence **Conj**[unction] will be found in the phrase and clause levels. At the phrase level, conjunctions may be used for modifiers as in *a rare but aggressive cancer,* or for phrases as in *the adults and the kids*; see Table 3.

Analyzed Text	*a*	*rare*	*but*	*aggressive*	*cancer +*
Phrasal CEs	Mod	Mod	Conj	Mod	Head

Analyzed Text	*the*	*adults +*	*and*	*the*	*kids +*
Phrasal CEs	Mod	Head	Conj	Mod	Head

Table 3. CE annotations of *a rare but aggressive cancer* and *the adults and the kids.*

At the clause level, I have conventionally put the CE label in Table 1 in the column for the first word in any multiword clause CE. For ease of visual appearance, I suppressed the cell borders for phrase CEs of multiword phrases. The CEs include **Pred**[icate] and **Arg**[ument], where 'argument'

is defined broadly as referring phrases denoting participants in the event (including so-called "adjuncts"). Arguments are not divided into subtypes (Subject, Object, Oblique, or alternative subtype classifications).

Complex predicates are somewhat problematic for a syntactic annotation scheme that divides a sentence into spans (continuous strings of words). The parts of complex predicates do not necessarily occur as contiguous strings. Thus they cannot generally be assumed to form a "predicate phrase". Nevertheless, in the great majority of cases, there is a complex predicate part which is clearly the primary predicate in the clause, and the other complex predicate parts are related to the primary predicate. We have chosen to distinguish the primary predicate with Pred; to label the other elements **CPP** (for 'complex predicate part'); and to treat the Pred and each CPP as CEs of the clause, rather than to group contiguous Pred and CPP units as a type of phrase. As with Mod and Arg, we do not distinguish different subtypes of parts of complex predicates.

The Clausal CE level is where phrasal "modifiers" are found, since recursion is excluded in the annotation. Hence there are also CE labels for **Gen**[itive] phrases and for adpositional and case-marked noun modifiers; since the latter two are grouped together as flags in recent typological terminology, we call the latter **FPM** for 'flagged phrase modifier'. Finally, a (coordinating) **Conj**[unction] linking the two clauses is a CE at the clausal level.

At the Sentential CE level, there are two clauses. Both are annotated **Main**, since the example in Table 1 is a coordinate construction of main clauses. Other CE labels are the standard trio of subordinate clauses, **Adv**[erbial], **Comp**[lement] and **Rel**[ative] clause. There also exist topic and focus phrases, sometimes analyzed as separate from clauses and hence sentence level CEs. We will label these as **Dtch** (for 'detached phrase'). Finally, there is Heine's category of **Int**[eractive] CE that occurs at the sentential construction "level".

Another example of CE annotation with an interrupting relative clause and genitive phrase, based on example (15), is given in Table 4.

Analyzed Text	*The*	*tree +*	*{that +*	*fell +*	*on*	*{our}*	*house}*	*also +*	*wrecked +*	*your +*	*car .*
Phrasal CEs	Mod	Head	Head	Head	Adp	Head	Head	Head	Head	Head	Head
Clausal CEs	Arg		Conj	Pred	Arg	Gen		CPP	Pred	Gen	Arg
Sentential CEs	Main		Rel								

Table 4. CE annotation of *The tree that fell on our house also wrecked your car.*

As with the first table of CE annotations, I have removed borders around empty cells when they are part of a multiword construction. However, because of the interrupting constructions, cell borders corresponding to the closing curly braces for the interrupting construction are retained, again for visual readability.

The CE labels proposed here, and the construction types they are found with, are given in Table 5. There are 14 labels for CE roles, and 5 labels for CE strategies.

CE tier	***Construction types***	***Label***	***Description***
Phrasal CEs ("lexical")	***roles played by words, including fixed MWEs [X^Y^Z]***	Head Mod Adm	head of phrase modifier [Dem, Num, Adj etc.] admodifier
	strategies	Adp Lnk Clf Idx Conj	adposition linker classifier index conjunction
Clausal CEs ("phrasal")	***roles played by phrases (which may be single words); or words (predicates, CPPs/"adverbs")***	Pred Arg CPP Gen FPM	predicate argument phrase [Sbj, Obj etc.] complex predicate part genitive phrase modifier flagged phrase modifier
	[*strategy*	Conj	conjunction**]**
Sentential CEs ("clausal")	***roles played by clauses (which may be single words); phrases ("detached NPs"); words (interactives)***	Main Adv Rel Comp Dtch Int	main clause adverbial clause relative clause complement clause detached phrase (topic, focus) interactive

Table 5. Summary of CE annotation labels and the syntactic levels they are found in.

3.2. Automatic CE annotation

The information needed to automatically generate CE labels is already largely available in (i) the flat syntax annotation introduced in §2, (ii) the gloss line with the interlinear morpheme translation, (iii) descriptive information about word class and (iv) and basic word order patterns. Word class and basic word order patterns is information that a descriptive linguist would document anyway in a lexicon and grammar and could be made available in a standard format. This information can be used to automatically generate the CE annotation via software developed for language description to a reasonable degree of accuracy. In this section, I will briefly summarize the sort of information that would contribute to the annotation of CE labels.

Phrases employ certain concepts which are expressed either as affixes that are glossed in the IMT or as closed-class function word categories which can be listed or linked from a lexical entry that already exists or is generated during text glossing. Table 6 lists these concepts and their options.

Concept	*CE label*	*Function word class*	*Bound morpheme*
reference	Head		number, gender, definiteness, case affixes
modification	Mod	linker, classifier	modifier index, attributive marker
admodification	Adm	admodifier	instensifier, downtoner, hedging morpheme
[words for strategies]	Adp, Lnk, Clf, Idx, Conj	adposition, linker, classifier, index, phrasal conjunction	case affix, linking affix, bound classifier, indexation

Table 6. Grammatical concepts occurring in phrase constructions.

Although the CE label Mod proposed in §3.1 does not distinguish subtypes of modifiers, such as Art[icle], Dem[onstrative], Num[eral], Adj[ective] and possibly other modifier types, this information could be derived from a basically semantic classification of modifier words in the lexicon.

In addition, word order patterns would indicate for instance that the admodifier in *a very large tree* in Table 2 modifies the adjective *large* because of the Adm-Adj order of English.

Clauses, like phrases, employ certain concepts that are expressed as affixes glossed in the IMT or closed-class function word categories that can be listed in a lexical entry. Table 7 lists such concepts and their expression. Adpositions, linkers and phrasal conjunctions are parts of phrases, but are actually used to indicate clause-level CEs. They are indicated by * in Table 7. Some morphosyntax is associated with a head on which the labeled CE is dependent (serving as an argument or a possessor). These are indicated by † in Table 7.

Concept	*CE label*	*Function word class*	*Bound morpheme*
predication	Pred		tense, aspect, modality, polarity affixes
participant, argument	Arg	adposition*	flag (case marking), indexation†
possession	Gen	adposition*, linker*	genitive flag/linker, indexation†
object modifier	FPM	adposition*, linker*	spatial/other flag
predicate part	CPP	auxiliary, manner adverb	flag, predicate inflections
joining	Conj	phrasal conjunction*	bound conjunction morpheme

*phrasal CE
†found on Head (predicate, or possessed head noun for Gen)

Table 7. Grammatical concepts occurring in clause constructions.

Although the CE label Arg proposed in §3.1 does not distinguish specific grammatical roles such as Sbj, Obj, Obl—or other classifications such as Erg, Abs or P[rimary]O[bject], S[econdary]O[bject]—the flags and indexes often associated with Arg and the Pred head of the clause can provide that information.

Basic word order patterns such as the order of subject, verb, object and oblique, the order of genitive and (head) noun (phrase), and the order of flagged phrase modifiers will help in

identifying genitive modifiers vs. argument phrases, and which referring phrase serves as the head of the genitive construction.

Sentences, like clauses and phrases, employ concepts expressed as affixes or closed-class function words. Table 8 lists the concepts and their expression. are parts of clauses, but are actually used to indicate sentential CEs.

Concept	*CE label*	*Function word class*	*Bound morpheme*
assertion	Main	coordinating conjunction*	predicate inflections†, deranked predicate (chaining form, switch-reference)†, absence of subordinating conjunctions
non-asserted event (adverbial)	Adv	adverbializer*	deranked predicate (converb)†
action as argument (complement)	Comp	complementizer*	deranked predicate (action nominal, infinitive, etc.)†
action as modifier (relative clause)	Rel	relativizer*	deranked predicate (participle)†
pragmatically defined referent	Det	topic/focus marker	topic/focus affix
interactive	Int	interjection, discourse marker ideophone, etc.	[reduplication for ideophone function]

*clausal CE
†occurs on Pred clause CE in sentence construction

Table 8. Grammatical concepts occurring in sentence constructions.

Here, some sentence CE (i.e. clause) types are identified by a clause CE contained in the clause, such as the different conjunctions. In other cases, the sentential CE types are identified by affixes on the predicate that heads the sentence CE (clause).

As with other automatic labeling algorithms, the results will not be perfect. Affixes, associated function word classes, and word order patterns are cues to the identity of basic syntactic constructions used to communicate information. Not all these cues are available in every language, although there is some evidence suggesting that the absence of some cues are made up by the presence of other cues (Shcherbakova et al. 2024). Some cues are polyfunctional: for example, the English subordinator *that* may introduce a relative clause (*the bicycle that I bought*) or a complement (*I told her that I bought a bicycle*).

The cues will provide more or less information for CE labeling depending on how they are annotated in IMTs and lexicons. For example, glossing English *that* as COMP (complementizer) vs. REL (relativizer) in a text, instead of SUBR (subordinator) or CNJ (conjunction) for both, will facilitate labeling the corresponding clauses as Comp vs. Rel. Finer-grained classification of function words in terms of their grammatical functions will also improve CE labeling. Also, the use of standardized abbreviations for affixes that reflect the comparative concepts they instantiate will allow for a general cross-linguistic algorithm for labeling CEs. It is our hope that manual correction of CEs will be minimal enough for automatic CE labeling to be useful, along with the

annotation of the syntactic boundary markers in §2. However, at this point this is a hope for the future.

4. Do we need recursion or embedding to represent syntactic structure?

The flat syntax annotation scheme described in §2 eschewed recursion for practical reasons. It simplified the annotation of the majority of phrases and clauses by treating them all as if they were at the same syntactic "level" (phrasal or clausal). This allows the use of boundary marker annotations that function in basically the same way as morphological boundaries in glossing words. Although a special notation for interrupting ("center-embedded") phrases and clauses was required to keep track of the two parts of the phrase or clause that is interrupted, this allows us to annotate all phrases at one "level" and all clauses at another "level" by using the same notational device that is used for infixes at the morphological level. This annotation provides a very simple syntactic structure with a minimum of notational means that is already familiar to descriptive linguists from morphological segmentation in interlinear glossing.

Having confined all syntactic annotation to three "levels", representing a sort of "concatenative syntax", we proposed a set of construction element (CE) labels that restores much of the structural information to the annotation, for users willing to add CE labels on the three levels. These CE labels resemble many of the Universal Dependencies labels for relations between words. (In some cases, labels are less specific, such as Mod covering the UD relations *amod, nummod*, and *det*, and Ref covering the UD relations *subj, obj, obl*.) However, the UD relations are relations between a dependent word and its head word, while the CE labels are roles, i.e. a relation between the construction role in question and the construction it plays a role in. In constructions with heads (likely a majority of construction tokens) the translation from the flat syntax notation to UD notation should be fairly straightforward.

However, the information typically represented by recursive syntax is less easy to reconstruct, due to the limitations of the flat syntactic annotation. The question I wish to address here, is if this limitation of the flat syntactic annotation is a bug—or is it perhaps a feature? Is there any reason to have recursion in the syntactic representation? I will suggest here that it is not.

4.1. "Main" vs. "subordinate" clauses

I will start with multiclause sentences, probably the most common type of construction that is represented by recursive or embedded syntactic structures. The point of including recursion is that a clause can be in an asymmetric relation with another clause which can in turn be in an asymmetric relation with yet another clause, and so on indefinitely in principle.

Constructions that invite representation by multiple embedding are rare. But another question to ask is: how are they encoded? I do not know of a language that morphosyntactically distinguishes multiple embedding from single embedding, or indicates depth of embedding. That is, I do not know of a language that indicates morphosyntactically, for example, that an adverbial clause is embedded under another adverbial clause, or another subordinate clause, rather than a main clause; or distinguishes single, double, or triple embedding.

We linguists, or we hearers, interpret a sentence as involving multiple "embedding"; but that is based on the semantic and/or pragmatic relations between the events expressed by the different clauses. As Dryer (2009) observes in a paper considering a flatter (but still recursive) syntactic structure in explaining word order patterns, hearers are primarily parsing a sentence for meaning.

They must figure that out based on the information in the utterance: the words and how they are combined. But they aren't given any clues from the syntactic structure of the sentence about relations between events, apart from word order (more on which later).

It does appear that there is morphosyntactic structure that indicates a single level of embedding of clauses, i.e. main clause vs. subordinate clause. Deranked verb forms—verb forms that overtly code their relation to the sentence—and subordinating conjunctions distinguish putatively "subordinate" clauses from main clauses, and sometimes there are differences in word order between main clauses and subordinate clauses.

However, the semantic relations between events do not inherently indicate a recursive, as opposed to a chaining relation. The same semantic relations—temporal, causal, conditional etc.—are found in coordinate constructions and in adverbial subordination constructions. English examples of some of the semantic relations are given in (22) (Croft 2022:466, from Table 15.1):

(22) a. He washed the car **before** driving to the party.
a′. He washed the car **and** drove to the party.
b. She went to bed **because** she was exhausted.
b′. She was exhausted **and (so)** she went to bed.
c. He got into the army **by** lying about his age.
c′. He lied about his age **and** got into the army.
d. **If** you do that, the terrorists have won…
d′. Murphy, you do that **and** the terrorists have won… [attested example]
e. **Although** John had no money, he went into this expensive restaurant.
e′. John had no money, **but** he went into this expensive restaurant. (König 1988:151)

Therefore one must look somewhere else than the semantic relations between events to capture the difference between main clauses and subordinate clauses.

Cristofaro (2003), following others, argues for a pragmatic or information-packaging distinction between main and subordinate clauses. The event expressed in main clauses is pragmatically asserted, while the event expressed in subordinate clauses is pragmatically non-asserted. This pragmatic contrast is then reflected in the morphosyntactic contrast between main clauses and subordinate clause constructions.

Except that sometimes it isn't. Coordinate clause constructions assert the events in each clause, and yet use a deranked strategy typical of "subordinate clauses" (Stassen 1985; Haspelmath 1995, 2004). Compare example (23), from Japanese, with its English translation (Yuasa and Sadock 2002:92; cf. Haspelmath 2004:34):

(23)

ojiisan-ga	yama-de	hatarai-**te**	obaasan-ga	mise-no
old.man-NOM	mountain-at	work-**CO**	old.woman-NOM	store-GEN

ban-o	shi-ta
sitting-ACC	do-**PST**

'The old man worked at the mountain, **and** the old woman tended the store.'

The phenomenon described as 'insubordination' (Evans 2007; Evans and Watanabe 2016) is another instance of the use of morphosyntactic structures associated with subordinate clauses to express an asserted event, or a speech act that is normally expressed by a main clause. For example, the imperative constructions illustrated below use the Spanish Subjunctive in (24), the English Gerund in (25) and the Russian Infinitive in (26) (Croft 2022:479):

(24) Díg-an-se-lo Ustedes.
tell:3SG.**SUBJ**-REFL-3SG.OBJ 2PL.FORMAL
'Tell them about it.'

(25) No smok**ing**.

(26) Ne kur-it'.
NEG smoke-**INF**
'No smoking.'

Matisoff (1972) describes a Lahu nominalizing particle *ve*, also used for subordination, that also occurs with main clauses. Example (27) illustrates both the subordination function and the main clause use (Matisoff 1972:245, gloss added [cf. Matisoff 1988]):

(27) là? tha ve šɨ̂ ɛ̀ qay ve
hand clap NOM make_noise go NOM
'The hand-clapping was boisterous.'

Conversely, certain constructions that are standardly associated with main clauses, such as the English inversion constructions in (28), do not normally occur in subordinate clauses, as in (29); but can occur in what is morphosyntactically a subordinate clause in the right syntactic and pragmatic context, as in (30) (the (a) examples are from Green 1976:383, the (b) examples from Lakoff 1984:472):

(28) a. Never before have prices been so high.
b. Here comes my bus.
(29) a. *Nixon regrets that never before have prices been so high.
b. *I'm leaving, if here comes my bus.
(30) a. I knew that never before have prices been so high.
b. I'm leaving, because here comes my bus.

Broadly, the purported main clause morphosyntactic structures in the (a) sentences are what Lakoff calls 'speech act constructions': 'constructions that are restricted in their use to expressing certain illocutionary forces that are specified as part of the grammar of English' (Lakoff 1984:473). However, in certain contexts, these illocutionary forces may be expressed in syntactically subordinate clauses, in the (c) sentences but not the (b) sentences. The contexts are fine-grained and difficult to characterize precisely (see Green 1976, Bolinger 1977 and Lakoff 1984 for qualifications). The main point for us, however, is that the morphosyntactic structure in question cannot be a sufficient indicator of main clause status.

All of these phenomena demonstrate that certain morphosyntactic strategies are not in themselves signals of syntactic main clause vs. subordinate clause, that is, clause recursion, or even clause embedding. They are at best fairly but not perfectly reliable indicators of certain asymmetric semantic relations or certain pragmatic statuses (assertion and other speech acts, vs. non-asserted propositions). There is no necessary reason to represent the distinction in syntax as one between a matrix clause and an embedded clause, i.e. one-step recursion.

In fact, there may not be a necessary reason to represent the semantic or pragmatic distinction itself by a recursive structure either. We already noted that the semantic relations between events may be expressed by coordinated main clauses—or simply sequence in discourse—or by main clause + subordinate clause. There is an asymmetric semantic relation between events—anterior-posterior time, cause-result, protasis-apodosis, and so on. There is also an asymmetric pragmatic relation between events, albeit one of contrast in pragmatic status (asserted vs. non-asserted), not a relation between those two different statuses. But just as asymmetric semantic relations between two events can be expressed by coordination of two main clauses, our semantic representation of an asymmetric relation between two events does not need to involve an embedding relation, such that one event (the embedded event) is **part of** the other event. Likewise, the pragmatic contrast between asserted and non-asserted events does not need to be represented in pragmatic structure such that the non-asserted event is **part of** the asserted event. The assumption that an asymmetric relation between entities (syntactic, semantic, or pragmatic) should or must be represented by a part-whole embedding relation can be called the **representing asymmetry by recursion assumption**. It is not a necessary assumption for syntactic representation; and it is not a necessary assumption even for semantic or pragmatic representation.

4.2. Interrupting units: why do they exist?

In §4.1 I argued that relations between syntactic units at the same "level", i.e. relations between clauses or relations between phrase, do not need to be represented in an asymmetric fashion such that one is syntactically recursively embedded in the other. However, it is true that one unit at a particular "level" may sequentially interrupt another unit as the same "level": a clause may interrupt another clause, or a phrase may interrupt another phrase (see §2.4).

There are two sets of grammatical constructions that commonly give rise to interruption, and are usually analyzed as involving syntactic embedding. At the clause "level", complex sentence constructions which are most likely to appear to involve clause embedding are complement clause constructions and relative clause constructions. A complement functions as the argument of the matrix clause predicate, and a relative clause functions as a modifier of an argument of the matrix clause predicate. Hence both appear to be parts of the matrix clause, although they are clauses themselves. Adverbial clauses are much less likely to interrupt their matrix clauses.

At the phrase "level", complex referring phrase constructions that appear to involve phrase embedding are genitive phrase modifiers and spatiotemporal phrase modifiers. In both cases, a referent is used to modify the referent of the matrix phrase. Functionally, in both cases the modifying referent is functioning as the anchor (Fraurud 1990; Hawkins 1991; Koptjevskaja-Tamm 2002) or reference point (Langacker 1993; Taylor 1996). Genitive modifiers are better described than spatiotemporal phrase modifiers, so our suggestions will apply largely to those.

The fact of sequential interruption has led syntacticians to conclude that recursive embedding is necessary to represent the fact that what precedes and follows the embedding structure belongs to a single syntactic unit (clause or phrase). Karlsson (2010:50) states that recursion is necessary

for a center-embedded (interrupted) structure; apparent recursion at the left or right edge can be converted to iteration. This is not a necessary conclusion for representing syntactic structure, as the representation of infixation in morphology shows (this is, however, contrary to the assumptions of formal language theory that Karlsson describes). We adopt the same representation for interrupting units in syntax.

Nevertheless, one might want to ask, why do speakers allow for interrupting units, and relatively commonly? After all, infixation is rare in morphology, whereas it is likely that even with the flat syntax analysis argued for in §4.1, probably all languages accommodate interruption. In this section, we suggest some reasons why interruption is found widely, but not that frequently in actual language use.

Interruption is basically an issue for processing. In spoken language, with rapid real-time sequential presentation and then decay of grammatical units, the processing load is great enough that interruption can be problematic. In writing, the reader can view strings of grammatical units at once, and so interrupted structures may be more frequent (cf. Karlsson 2010:64).

Karlsson (2007, 2009, 2010) presents evidence that interruption, and even asymmetric relations between contiguous but not interrupting units, is quite limited in depth of embedding, more common in written than in spoken language, and evolved later than coordination in emerging written languages (Karlsson 2009). We focus on center embedding, although Karlsson provides evidence that recursion at left and right edges is only slightly less constrained. Karlsson's evidence is based on texts from several "Standard Average European" languages. He concludes that 'no evidence for nested [center-embedded] syntactic recursion of degrees greater than 3 is at hand, neither on clause level nor on phrase level' (Karlsson 2010:63). Phrasal center-embedding has been less well studied, but even in written language, a depth of 2 is 'utterly rare' (Karlsson 2010:58).

Multiple clausal center-embedding in spoken language occurs only extremely rarely at a depth of 2 (Karlsson 2010:53): 'only one cycle of clausal center-embedding is normally to be found in conversation, the most basic and evolutionarily primary form of language' (Karlsson 2019:315). That is, in the latter case at least, interruptions of interruptions at the clause level are close to nonexistent in spoken language.

Interruption of syntactic units, and even multiple asymmetric relations in a string of syntactic units, is thus relatively uncommon. Why is this possible? That is, what allows speakers to express what may be complex asymmetric functional relations between semantic or pragmatic components such as events or referents in a sequential manner without many interruptions? I suggest here and in the following section that grammatical constructions have evolved in such a way to at least allow for a speaker to avoid interrupting syntactic units with other syntactic units at the same "level". In §4.2.1, I suggest that common word order patterns frequently allow for noninterrupting sequencing of syntactic units at the same "level". In §4.2.2, I suggest that grammaticalization processes lead to the reanalysis of a syntactic unit that is functionally dependent on another syntactic unit so that it belongs to a lower syntactic "level".

4.2.1. Word order "conspiracies" to minimize interruption

One consequence of a flat syntactic model is that word order plays a greater role in syntactic processing than embedding. Most languages have a basic word order for different kinds of clausal and phrasal units that are related to other CEs at the same "level", e.g. Verb-Complement order, Genitive-Noun order, Relative Clause-Noun order, or Subject-Verb-Object order. Many if not all

languages also allow for alternative word orders for some of these units, even when they express the same semantic or pragmatic relation between these units. Interruptions only occur under some circumstances given common word order patterns. Alternative word orders, and certain strategies for constructions that are found in some languages, provide means to combine clauses or phrases without one interrupting the other. In this section I describe some general word order patterns that do or do not lead to one phrase or clause interrupting another.

We begin with phrases. In example (18) in §2.4, the English possessive phrase *my* interrupts the referring phrase with a preposition: *on {my} house*. This occurs because the English pronominal possessive, as well as the nonpronominal *-'s* possessive, precedes the noun, and so do English adpositions (i.e., prepositions). This pattern, which we can summarize as GenN and Prep, following the common abbreviations used in word order studies, is actually a rare word order combination. The overwhelming pattern is for languages to have either GenN order and postpositions, or NGen order and prepositions. These combine as in example (31), without one phrase interrupting the other:

(31) a. PrepN and NGen: + Prep N + Gen +
b. NPostp and GenN: + Gen + N Postp +

In fact, English also has another genitive construction, the *of* phrase (*the surface of the table; a book of John's*), which has NGen order. The other major phrasal modifier construction of nouns in English, an oblique phrase with a preposition that generally expresses spatiotemporal relations (annotated FPM in §3), also follows the noun in English. I am not aware of crosslinguistic studies of the word order patterns of non-genitive referring phrases functioning as modifiers of referents. However, the genitive construction pattern suggests that genitive phrases interrupting adpositional phrases is crosslinguistically uncommon. The anomalous English pattern illustrated in (18) is discussed further in §4.2.2.

The situation with clauses is more complex. As noted in §4.2, adverbial clauses are usually peripheral to their matrix clause, and rarely interrupt it. Complement and relative clauses serve as arguments of the matrix predicate, or modifiers of arguments, respectively. The default ordering places such clauses in the same position as the corresponding phrasal arguments. This can lead to the complement/relative clause interrupting its matrix clause under some circumstances. But crosslinguistic word order patterns, text frequency patterns, and the existence of alternative strategies suggest that this may not be that common a phenomenon.

The two most common word order patterns for arguments and their predicate are Subject-Object-Verb and Subject-Verb-Object, usually abbreviated SOV and SVO. Assuming that complement clauses occur in the same position as the corresponding phrasal arguments, the complement clause sometimes interrupts the matrix clause, and sometimes does not:[13]

(31) a. SVO, subject complement: # Comp # V + Obj #
b. SVO, object complement: # Sbj + V # Comp #
c. SOV, subject complement: # Comp # Obj + V #
d. SOV, object complement: # Sbj {Comp} V #

[13] This discussion does not address the occurrence of CEs of the matrix clause other than subject, object and verb/predicate. For many of the non-argument, non-predicate CEs of the clause, not enough is known about their typological word order behavior. It is my impression that the presence of other CEs than arguments and predicate does not substantially increase the occurrence of interrupting/center-embedding of clauses and phrases.

Only in (31d) does the complement clause interrupt the matrix clause. However, there are two text frequency patterns that suggest that complement interruption may not be that frequent. First, transitive clauses in general are no more common, and perhaps less common, than intransitive clauses. Object complements are not found in intransitive clauses by definition.

Second, there is an asymmetry in the expression of transitive subject referents—the A grammatical role, in typological parlance—in comparison to transitive object referents (P) or intransitive subject referents (S). A referents are very frequently highly accessible pragmatically. In the great majority of languages this means that there is no overt transitive subject phrase, because highly accessible referents are expressed only by indexation on the predicate or not at all (null anaphora). This pattern has been described as Preferred Argument Structure (DuBois 1987; DuBois et al. 2003). In other words, interruption of a matrix clause by a complement clause occurs only in transitive matrix clauses with an overt subject phrase—a textually rare pattern. (And this happens only in SOV languages.)

The situation with relative clauses is more complicated. The vast majority of relative clauses, crosslinguistically, are externally headed (other relative clause strategies are discussed below). The external head is an overt argument of the matrix clause. The relative clause may occur either before or after the external head; however, preposed relative clauses are crosslinguistically very rare in SVO languages. Leaving this rare case aside, the pattern of interrupting relative clauses is given in (32)-(33):

(32) a. SVO & NRel, relative clause on subject phrase: # Sbj {Rel} V + Obj #
b. SVO & NRel, relative clause on object phrase: # Sbj V Obj # Rel #

(33) a. SOV & RelN, relative clause on subject phrase: # Rel # Sbj + Obj + V #
b. SOV & NRel, relative clause on subject phrase: # Sbj {Rel} Obj + V #
c. SOV & RelN, relative clause on object phrase: # Sbj {Rel} Obj + V #
d. SOV & NRel, relative clause on object phrase: # Sbj + Obj {Rel} V #

There are more situations in which a relative clause may interrupt its matrix clause than for a complement clause to interrupt its matrix clause. In particular, there are many more situations in which a relative clause could interrupt its matrix clause in SOV languages. However, one of those cases, (33c), would only cause interruption if the transitive subject/A referent is expressed by an overt phrase, which is less common.

Nevertheless, there are alternative strategies for expressing relative clauses that do not interrupt the main clause. In some languages such as English, a relative clause on a subject phrase can be extraposed, as in example (34):

(34) *The tree + was + dead # that + fell + on {my} house.*

English also allows postposing of complement clauses, albeit with a different construction using the pronoun *it:*

(35) a. *That + he + resigned # isn't + surprising.*
b. *It + isn't + surprising # that + he + resigned.*

Although neither (35a) nor (35b) interrupt the matrix clause, postposing a relative or complement clause can make a bigger difference in SOV languages, where both argument positions are normally before the verb or predicate. Some SOV languages are "nonrigid", that is, some CEs of the clause may follow the verb (Greenberg 1966:79-80). In nonrigid SOV languages, if complement and relative clauses were also postposed, then they would not interrupt a matrix clause containing a predicate and an overt argument phrase (as well as other CEs of the matrix clause). In this context, it is also worth noting that the direct report strategy for utterance complement-taking predicates (which is extended to propositional attitude predicates in some languages), also postposes or sometimes preposes the utterance complement to the matrix clause.

There are also other, albeit crosslinguistically rare, strategies for relative clauses that reduce the likelihood that a relative clause will interrupt the matrix clause. Adjoined relative clauses express the relative clause similarly to an adverbial clause. Correlative relative clauses juxtapose the relative clause to the matrix clause. These two strategies lead to the relative clause not interrupting its matrix clause.

Finally, internally-headed relative clauses do not have an externally expressed head; the head referent is expressed as an argument phrase inside the relative clause. In both SOV and SVO languages, an internally headed relative clause on the subject phrase does not interrupt the matrix clause. In SOV languages, an internally headed relative clause on the object phrase does not interrupt the matrix clause if the matrix subject referent is not expressed as an argument phrase—the common case, according to Preferred Argument Structure. Given that internally headed relative clauses minimizes interruption in three out of four cases for SOV languages, it might not be surprising that internally relative clauses, while crosslinguistically rare, are overwhelmingly more frequent in SOV languages.

Internally headed relative clauses are very rare in SVO languages. Likewise, RelN word order is very rare in SVO languages. Thus, (36), a postposed relative clause on a subject argument phrase in an SVO language, will almost always interrupt the matrix clause:

(36) *The man {who's + picking + pears} comes + down + from the tree.*

In fact, as has frequently been noted, postposed relative clauses in English are always broken into a separate intonation unit, while internal relative clauses are always kept in the same intonation as their head (Croft 1995:847-48 and references cited therein; X,Y is speaker number, intonation unit number):

(37) 10,29 *a--nd put them in a couple of [.25] barrels,*
10,30 *that he's got down there.*

(38) 7,78 [1.35 + [1.35 *Meanwhile...*] *the man who's picking pears,*
7,79 [.35] *comes down from the tree.*

The regularity of this pattern, in English at least, indicates that although the relative clause on the subject phrase in (38) is morphosyntactically externally-headed, the speaker's prosodic production suggests that the subject phrase is construed as part of the relative clause, despite the conventional syntactic analysis of the construction.[14]

[14] It is possible to interpret the prosody of (38) as capturing the combination of the external head and the relative clause as the subject phrase. However, this interpretation would not explain why this prosodic pattern always occurs

These general word order patterns, and in some cases alternative strategies, for expressing clauses and phrases that are traditionally analyzed syntactically as embedded strongly suggest that interruption of one phrase or clause by another is a disfavored strategy, at least in spoken language. In some cases, it may even suggest a different syntactic analysis of the construction, as with example (38). These conclusions can, and should, be tested further by text counts that would more definitely confirm that interruptions are minimized, but this task is beyond the scope of this chapter.

4.2.2. Interrupting units: "downgrading" an interrupting unit to a lexical dependent or even an affix

The typological and text frequency phenomena described in §4.2.1 indicate that phrases or clauses that interrupt other phrases and clause may be relatively rare in actual language use, although they undoubtedly occur (hence the need for the { } annotation for syntactic boundaries). But another factor that may reduce interrupting units is a reanalysis of the interrupting unit itself, in the course of grammaticalization processes that affect many such structures.

The main problem with interrupting units is when they consist of more than multiple words. In that case, it is clear that they form a phrase, or a clause, and that creates a potential jumble of two (or more) syntactic units of the same "size". But in fact, very often the interrupting unit is a single word, or at most a single content word combined with a word that is part of the interrupting construction's strategy, such as a relativizer or a linker. This is not necessarily an accident or random factor in production. The single-word "phrase" or "clause" may actually be grammaticalizing into a form that is better (re)analyzed to belong at the lexical "level". Additional evidence of this reanalysis is that the form further grammaticalizes into a bound affix which is no longer a syntactic unit.

A good example of this involves genitive phrases. As noted above, English has an anomalous word order of GenN and Prep, leading to the preposed possessor interrupting the referring phrase, as in + *on* {*our*} *house* +. This pattern is found for pronominal possessors across many European languages, even those where nonpronominal possessors are expressed in a postposed flagged phrase (i.e. using adpositions or case affixes). In functional terms, pronominal possessors are like prototypical phrases: they express referents. Hence our analysis as a phrase, and in the English example, a phrase that interrupts the prepositional phrase it modifies.

But pronominal possessors are almost always single words. They sometimes are expressed using modifier strategies typical of lexical modifiers, such as the nonperson indexation typical of adjectives and demonstratives, as in Russian: compare *mo-**ja** kniga* 'my book [FSG]' to *krasna-**ja** kniga* 'red book [FSG]'. In many languages, pronominal possessors have grammaticalized to person indexation affixes on the head noun denoting the possessed item. In languages such as English or Russian, it might be better to consider at least possessive pronouns to be lexical modifiers rather than phrasal modifiers, in which case the annotation would be, for instance, + *on our house* +.

At the clausal level, things are a bit more complicated but similar. Property concepts in many languages recruit the relative clause strategy found with action concepts as modifiers, with verbal inflections and perhaps a relativizer. However, most property concepts do not have any arguments apart the role filled by the head noun referent that the property concept modifies. As a result, property concept "relative clauses" usually consist of just the property concept word (possibly with

with relative clauses on subjects, but never with relative clauses on objects (or obliques that occur at the end of a clause).

verbal inflections), and perhaps also a relativizer serving as a linking strategy to the head noun. In such languages, it might be better to consider property concept modifiers as lexical modifiers, despite the relative clause trappings that they bear.

Complement clauses are found with a wide variety of matrix clause predicates (complement-taking predicates or CTPs). These include utterance ('say'), propositional attitude ('believe'), perception ('see', 'hear'), desiderative ('want'), manipulative ('make', 'let'), modal, aspectual and negation CTPs. Semantically, the CTP denotes anything from an event independent of the event expressed in the complement clause (utterance, propositional attitude) through more or less connected events (perception, manipulative) to semantic properties of events (modal, aspectual, negation).

The argument structures of the matrix clause and the complement clause range from being independent or partly independent (utterance, propositional attitude, perception) to being partly shared (manipulative) or fully shared (modal, aspect, negation). In some cases, it is not always clear which clause the argument phrase belongs to, or whether there are even two clauses or just one.

Finally, even the CTPs that denote independent or partly independent events may grammaticalize to a lexical form that does not, or no longer, looks entirely like an independent clause predicate (auxiliaries) to an invariant form and ultimately an affix on the complement clause predicate, as indicated in (39):

(39) a. utterance CTPs > quotatives
b. propositional attitude CTPs > hearsay, inferential evidentials and epistemic modals
c. perception CTPs > sensory evidentials
d. desiderative CTPs > desideratives
e. manipulative CTPs > causatives
f. modal, aspectual, negation CTPs > modal, aspectual, negative markers

Again, it is plausible to analyze constructions at a certain intermediate stage in this grammaticalization process as no longer consisting of a two-clause complement clause construction. Instead, the CTP would be analyzed as a complex predicate part (CPP), such as an auxiliary, instead of a predicate heading a separate "matrix" clause. This is more justifiable when there is a single argument structure (or some sort of merged argument structure), and the "former" matrix clause lacks other CEs than the CTP.

5. Conclusion

In this chapter, I propose a practical syntactic annotation scheme, consisting of four types of boundary markers in addition to word boundary (space) and sentence boundary (period) already used in glossed texts: + for phrases, # for clauses, { } for phrases or clauses that interrupt other phrases or clauses, and ^ for fixed contiguous multiword expressions. In addition, I proposed a small general set of labels for constructional roles for construction elements or CEs, for the syntactic units defined by the boundary markers. These are optional and it is hoped that annotating the roles can be automated to a great extent, given a glossed text including the syntactic boundary markers proposed here.

In order to make the syntactic annotation simple enough, it simplifies syntactic structure to a point beyond most syntactic analyses of constructions, even those in a construction grammar or

descriptive linguistic tradition, or in dependency-based syntactic annotation, that generally eschews the highly complex phrase-structure trees of generative grammar. In particular, it does without recursion of syntactic units, which allows us to use a fixed set of syntactic "levels" (sentence, clause, phrase, word, plus the sublexical morphological analysis of traditional text glossing). This simplification was made for practical reasons. But it raises an interesting question: how much complex syntactic structure is really justifiable? The last section of this chapter suggests that perhaps linguists, even those of us of a functional bent, are still asking syntax to do too much work.

References

Bolinger, Dwight. 1977. Another glance at main clause phenomena. *Language* 53.511-519.

Comrie, Bernard. 1989. *Language universals and linguistic typology* (2nd ed.). Chicago: University of Chicago Press.

Cristofaro, Sonia. 2003. *Subordination*. Oxford: Oxford University Press.

Croft, William. 1995. Intonation units and grammatical structure. *Linguistics* 33.839-882.

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

Croft, William. 2022. *Morphosyntax: constructions of the world's languages.* Cambridge: Cambridge University Press.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre and Daniel Zeman. Universal Dependencies. *Computational Linguistics* 47(2):255-308.

Dixon, R. M. W. and Alexandra Aikhenvald. 2003. *Word: a cross-linguistic typology.* Oxford: Oxford University Press.

Dryer, Matthew S. 2009. The branching direction theory of word order correlations revisited. *Universals of language today*, ed. Sergio Scalise, Elisabetta Magni and Antonietta Bisetto, 185-207. Berlin: Springer.

DuBois, John W. 1987. The discourse basis of ergativity. *Language* 64.805-855.

DuBois, John W., Lorraine E. Kumpf and William J. Ashby (eds.). 2003. *Preferred argument structure: grammar as architecture for function.* Amsterdam: John Benjamins.

Evans, Nicholas and Honoré Watanabe (eds.). 2016. *Insubordination.* (Typological Studies in Language, 115.) Amsterdam: John Benjamins.

Fortescue, Michael. 1984. *West Greenlandic*. Croom Helm Descriptive Grammars. Dover, New Hampshire: Croom Helm.

Fraurud, Kari. 1990. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics* 7.397-433.

Graczyk, Randolph. 2007. *A grammar of Crow.* Lincoln: University of Nebraska Press.

Green, Georgia. 1976. Main clause phenomena in subordinate clauses. *Language* 52.382-97.

Greenberg, Joseph H. 1966. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of grammar*, ed. Joseph H. Greenberg, 2nd edition, 73-113. Cambridge, Mass: MIT Press.

Haspelmath, Martin. 1993. *A grammar of Lezgian.* Berlin: Mouton de Gruyter.

Haspelmath, Martin. 1995. The converb as a cross-linguistically valid category. *Converbs in cross-linguistic perspective: structure and meaning of adverbial verb forms—adverbials participles, gerunds*, ed. Ekkehard König and Martin Haspelmath, 1-55. Berlin: Mouton de Gruyter.

Haspelmath, Martin. 2004. Coordinating constructions: an overview. *Coordinating constructions*, ed. Martin Haspelmath, 3-39. Amsterdam: John Benjamins.
Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45.31-80.
Hawkins, John A. 1991. On (in)definite articles: implicatures and (un)grammaticality prediction. *Journal of Linguistics* 27.405-42.
Heine, Bernd. 2023. *The grammar of interactives.* Oxford: Oxford University Press.
Karlsson, Fred. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics* 43.365-92.
Karlsson, Fred. 2009. Origin and maintenance of clausal embedding complexity. *Language complexity as an evolving variable,* ed. Geoffrey Sampson, David Gil and Peter Trudgill, 192-202. Oxford: Oxford University Press.
Karlsson, Fred. 2010. Syntactic recursion and iteration. *Recursion and human language,* ed. Harry van der Hulst, 43-67. Berlin: de Gruyter.
Karlsson, Fred. 2019. Multiple center-embedding in spoken English. *Tokens of meaning: papers in honor of Lauri Karttunen*, ed. Cleo Condoravdi and Tracy Holloway King, 309-17. Stanford: CSLI Publications.
Koptjevskaja-Tamm, Maria. 2002. Adnominal possession in the European languages: form and function. *Sprachtypologie und Universalienforschung* 55.141-72.
Lakoff, George. 1984. Performative subordinate clauses. *Proceedings of the Tenth Annual Meeting of the Berkeley Linguistics Society*, ed. Claudia Brugman et al., 472-80. Berkeley: Berkeley Linguistics Society.
Langacker, Ronald W. 1993. Reference point constructions. *Cognitive Linguistics* 4.1-38.
Matisoff, James A. 1972. Lahu nominalization, relativization and genitivization. *Syntax and semantics*, vol. 1, ed. John Kimball, 237-57. New York: Academic Press.
Matisoff, James A. 1988. *The dictionary of Lahu.* (University of California Publications in Linguistics, 111.) Berkeley and Los Angeles: University of California Press.
Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: an evergrowing multilingual treebank collection. *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 1659-1666. European Language Resources Association.
Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: an evergrowing multilingual treebank collection. *Proceedings of the 12th International Conference on Language Resources and Evaluation*, 4034-4043. European Language Resources Association.
Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70:491-538.
Shcherbakova, Olena, Damián E. Blasi, Volker Gast, Hedvig Skirgård, Russell D. Gray, and Simon J. Greenhill. 2024. The evolutionary dynamic of how languages signal who does what to whom. *Nature Scientific Reports* 14:7259.
Siewierska, Anna. 1984. Phrasal discontinuity in Polish. *Australian Journal of Linguistics* 4.57-71.
Stassen, Leon. 1985. *Comparison and universal grammar.* Oxford: Basil Blackwell.
Taylor, John R. 1996. *Possessives in English: an exploration in cognitive grammar.* Oxford: Oxford University Press.

Van Gysel, Jens E. L., Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a Uniform Meaning Representation for natural language processing. *Künstliche Intelligenz,* https://link.springer.com/article/10.1007/s13218-021-00722-w

Zingler, Tim. 2020. Wordhood issues: typology and grammaticalization. Ph.D. dissertation, University of New Mexico.