Chunk Stylebook

Steven Abney

1996

1 Introduction

Roughly speaking, a *chunk* is the non-recursive core of an intra-clausal constituent, extending from the beginning of the constituent to its head, but not including post-head dependents. A sample text is provided in the last section of this paper to give an intuitive idea of what we have in mind.

In marking chunks, we are interested only in their category and start and end points. Evaluation consists in marking concordance lines of corpus positions i with the category and end-point of the chunk (if any) that begins at i. For example:

41/1087:	with " significant " business	stemming]**vgx** from we
19/530:	The	SEC will probably vot
25/701:	arch and development , Mr. Lane	said]**vx** . // The rul
47/1233:	Co. , a maker of biotechnology	instrumentation and e
63/1746:	disturb or challenge a listener	. // And so it went t
21/583:	, that the proposed changes "	would substantially i
22/605:	" What the investors who	oppose]**vx** the propos
36/966:	Both	funds are expected to
62/1697:	nd movement from Saint-Saens 's	Sonata]**nx** for Clarin
22/611:	the proposed changes object to	most is the effect th
17/453:	The changes were proposed in	an effort]**nx** to stre

Chunk categories are:

NX	noun chunk
VX	verb chunk
INF	infinitive chunk
VGX	present participle or gerund chunk
VNX	past participle chunk
AX	adjective chunk
RX	adverb chunk

1.1 Maximal Chunks

A maximal chunk is a chunk that is contained in no other chunk. For example, in *[*$_{\text{NX}}$ *the [*$_{\text{AX}}$ *new] building]*, the NX is maximal, and the AX is not. We mark only maximal chunks—only the NX, in this case. "Maximal chunk" should usually be understood when we write "chunk".

It should be noted that (maximal) chunks do not necessarily partition the sentence—there may be material that does not belong to any chunk. For example, prepositions, coordinators, subordinators, adverbs, and punctuation often are not part of any chunk.

1.2 Chunks Syntactically Defined

Chunks are defined strictly syntactically, not semantically, functionally, lexically, etc. There is a relation between chunks and "tightly bonding" phrases, where "tightly bonding" includes such notions as prosodic phrasing, distributional co-occurence (e.g., mutual information), lexical selection, lexical unpredictability, collocation, semantic constituency, semantic unpredictability, etc. *However* "tightly bonding" does *not* define chunks. Here are some examples of tightly bonding phrases that are not chunks:

in [$_{\text{NX}}$ spite] of [$_{\text{NX}}$ his objections]
[$_{\text{VGX}}$ according] to [$_{\text{NX}}$ our information]
[$_{\text{NX}}$ she] [$_{\text{VX}}$ looked] up [$_{\text{NX}}$ the answer]
apart from [$_{\text{NX}}$ my good friend] and [$_{\text{NX}}$ colleague]
[$_{\text{NX}}$ it] [$_{\text{VX}}$ took] [$_{\text{NX}}$ place]

These examples will be discussed below.

2 Chunk Definitions by Category

We begin with a thumbnail sketch of English syntax. The kernel of a sentence is the tensed verb group, containing a main verb and its auxiliaries. Noun phrases and prepositional phrases are the commonest verb arguments; noun phrases can also appear adverbially (*last week*) or as (part of) measure phrases (*two weeks later*). Noun phrases constitute chunks, but prepositional phrases are considered to belong to a level above the chunk level and are not marked as chunks. Measure phrases are considered chunks, but they are usually part of a larger chunk, hence not marked, since only maximal chunks are marked.

Adjectives have four main functions: prenominal modifiers (*a big dog*), predicate adjectives (*the dog is big*), postnominal modifiers (*a dog as big as a house*) and secondary predicates (*as big as it is, you'd think that dog couldn't run so fast*). Only postnominal adjectives and secondary predicates are marked AX—prenominal modifiers and predicate adjectives are included in the associated NX/VX, hence not maximal.

Participles and gerunds are verb forms that function like adjectives or nouns, respectively. Participles have either the suffix *–ing* or *–(e)d/(e)n;* gerunds always end in *–ing.* Regardless of their function, *–ing* forms are marked VGX when maximal, and *–(e)d/(e)n* forms are marked VNX.

"Adverb" is a catch-all class. To avoid error-prone but rather useless decisions about idiosyncratic adverbs, only multi-work adverbial phrases are marked as chunks.

Coordination is a special process that introduces special difficulties. In general, coordinators separate chunks unless they are "trapped" inside a left branch of a chunk. For example:

> [NX we] [VX lack] [NX the ways] and [NX means] [INF to ...]
> [NX the Ways and Means Committee]

Coordination is discussed in more detail in section 3.4.

2.1 NX

An NX extends from the beginning of the noun phrase to the head noun.

- An NX begins *after* any possessor phrases. E.g.: *[NX John] 's [NX hat].* These pieces are combined at a level above the chunk level.
- Possessive pronouns are not considered to be possessor phrases: *[NX his hat].*
- Temporal noun phrases and, in general, noun phrases functioning adverbially, are NX. E.g.: *[NX he] [VX left] [NX last Monday].*
- NX includes pronouns and proper names.
- The same rules apply to names with complex internal syntactic structure as to common noun phrases.

  > [NX the City] of [NX New York]
  > [NX New York] 's [NX Finest]
  > [NX Chicago] 's [NX Goodman Theatre]
  > [NX Standard] & [NX Poor] 's [NX 500-Stock Index]

- Abbreviatory signs like '$', '%', 'mph' are treated as nouns: *[NX almost $ 4,000], [VX rose] [NX 30 %] to [NX 93 ppm].*
- Quantifiers functioning as verb arguments are treated as NX's: *[NX all] [VX is not lost], [NX some] [VX would say], [NX he] [VX did n't say] [NX much].*

- Quantifiers, comparatives, and superlatives appearing with determiners or PP complements are treated as NX's, even when they function adverbially: *[NX most] of [NX the time], [VX like] [NX it] [NX the most], [VX like] [NX it] most, [VX like] [NX it] [NX a lot], [VX don't like] [NX it] much, [VX like] [NX it] [NX best] of [NX all], [VX like] [NX it] best.*

- Predeterminers are included in NX only if they would otherwise constitute NX's of their own, e.g. *[NX all the men] [VX left]*, *[NX all] [VX left].*

  - Examples of predeterminers are: *[NX all the men], [NX both the men], [NX such a fuss], [NX half a mind], [NX many a day], [NX almost all the children].*
  - *Such* is something of a grey case, since *?such was made* is not very good, but *[NX such] as [VX were left] [VX were stale]* is fine, so we'll include it generally as a predeterminer.
  - Adverbs premodifying noun phrases are not treated as predeterminers. For example: *quite [NX a fuss], rather [NX a mess], not [NX a peep], only [NX the good], even [NX the best], just [NX the thing].* None of these appear as noun phrases on their own.
  - We also distinguish predeterminers from precoordinators:

    [NX both the men] [VX left]
    both [NX the men] and [NX the women] left

    Both the men and the women is actually ambiguous—one must decide from context whether there are exactly two men and an indefinite number of women, or simply two groups, one of men and one of women.

- Partial noun chunks, e.g., coordinands with ellipses, are marked NX: *both [NX the old] and [NX the new styles].*

2.2 VX

VX includes all modals, auxiliary verbs, and medial adverbs, but ends at the head verb or predicate adjective. E.g.: *[NX John] [VX certainly screwed] up [NX that time].*

- Only *do, have,* and *be* are auxiliary verbs: *[VX help] [VX pack], [VX begin] [VGX packing].*

- Verb particles are not included in VX. Of course it is important to recognize verb-particle constructions, but that job is not part of recognizing chunks.

- Light nouns and pieces of V-N idioms are not included in VX: *[VX took] [NX place], [VX take] [NX advantage] of [NX it].*

- Prepositional phrases are not included in VX unless they are trapped between other pieces:

  [NX John] of [NX course] [VX needs] [NX help]
  [NX John] [VX will of course need] [NX help]
  [NX the bill] [VX would , in effect , legislate] ...
  [NX she] [VX had , after all , sung] ...

- Predicate adjectives (but not predicate nominals) are included in VX: *[VX is fun], [VX is interesting], [VX is still difficult]*

- Copulas are treated as main verbs if they take something other than a participle or predicate adjective:

  [NX John] [VX is] [NX my brother]
  [NX John] [VX is] about [INF to find] out
  [NX my hope] [VX is] that [NX it] [VX will snow]

- Fronted auxiliaries constitute separate VX's:

  [NX who] [VX did] [NX you] [VX see] ?

2.3 INF

INF phrases are infinitive chunks starting with *to*. Bare infinitives (without *to*) are VX, not INF. As with verb chunks, medial adverbs immediately preceding *to* are included in INF.

2.4 VGX

Present participle/gerund chunk, head verb ending in *–ing*: *[NX the man] [VGX washing] [NX the car] [VX is] [NX John], [VGX washing] [NX cars] [VX is fun], [VX couldn't stop] [VGX drinking].*

- When part of a tensed verb group, VGX is not maximal, so not marked: *[NX John] [VX is washing] [NX the car].*

- VGX includes gerunds functioning as noun phrases.

  [VGX flying] [NX planes] [VX is dangerous]
  [NX flying planes] [VX are dangerous]
  between [VGX writing] [NX letters] and [VGX mailing] [NX them]

  (Note that the prenominal participle *flying* is not a maximal chunk.)

- Adjectives in *–ing* are *not* participles: *[AX very interesting]* vs. *[VGX interesting] [NX them] [VX is difficult].*

2.5 VNX

Past participle chunk, head verb ending in *–(e)n/(e)d.*

- Unlike with VGX, it is too difficult to make an adjective vs. participle distinction where past participles are concerned, so all (postnominal and secondary-predicate) adjectives that are morphologically past participles are marked VNX: *[*$_{\text{VNX}}$ *very tired], [*$_{\text{VNX}}$ *closed] for [*$_{\text{NX}}$ *the season], [*$_{\text{VNX}}$ *not interested], [*$_{\text{VNX}}$ *already so stacked] against [*$_{\text{NX}}$ *the little guy].*

2.6 AX

AX are adjective chunks, beginning with any premodifying adverbs and intensifiers and ending at the head adjective: *[*$_{\text{AX}}$ *completely silent], [*$_{\text{AX}}$ *as quiet] as [*$_{\text{NX}}$ *a mouse].*

- Prenominal adjectives and predicate adjectives do not constitute maximal chunks, so they are not marked. AX includes postnominal adjective phrases and secondary predicates that modify a noun they are not adjacent to.
- There are some quantificational and temporal adverbs like *only, even, not, just, already* that are not obviously inside the adjective phrase, since they behave in many ways like sentential adverbs. However, we will consider them to belong to the the adjective phrase: *[*$_{\text{AX}}$ *already so sad], [*$_{\text{AX}}$ *not even as difficult] as ...*. This also applies to VGX and VNX: *[*$_{\text{VNX}}$ *already so stacked].*
- However, coordinators and precoordinators are *not* included in AX: *[*$_{\text{AX}}$ *not difficult], both [*$_{\text{AX}}$ *difficult] and [*$_{\text{AX}}$ *dangerous], not [*$_{\text{RX}}$ *so much] [*$_{\text{AX}}$ *difficult] as [*$_{\text{AX}}$ *tedious].* This also applies to VGX and VNX.

2.7 RX

Multi-word adverb phrases.

- RX includes only *multi-word* adverb phrases. Particularly, it does include adverbs modified by other adverbs or intensifiers, and it does include adverbs modified by measure phrases: *[*$_{\text{VX}}$ *ran] [*$_{\text{RX}}$ *very quickly], [*$_{\text{VX}}$ *ran] quickly, [*$_{\text{RX}}$ *three weeks later], [*$_{\text{RX}}$ *hardly even] [*$_{\text{NX}}$ *his mother].*
- Nouns used adverbially are considered NX, not RX. E.g.: *[*$_{\text{NX}}$ *yesterday], [*$_{\text{NX}}$ *last week].*
- Existential *there,* as in *there was a unicorn in the garden,* is considered a pronoun, hence NX, not RX.

- RX does not include particles, connectives, interjections, subordinators, or anything else in the general class of sentential "grit".
- Verb particles and stranded prepositions are not RX.
- Prepositions used adverbially, without complements, are considered to be adverbs, and are RX if modified, e.g. *[*$_{\text{RX}}$ *three miles across].* But: *[*$_{\text{NX}}$ *three miles] down [*$_{\text{NX}}$ *the road].*
- The word "ago" is considered an adverb, and since it is always (to my knowledge) modified by a measure phrase, it is always RX: *[*$_{\text{RX}}$ *two years ago].*
- Sentence connectives and coordinators, even if multi-word like *then again, even though, but still, but also, and also,* etc., are *not* RX.
- Interjections, e.g. *yes, no, maybe, huh, uh, whoa, gee, golly,* etc., are not RX.

3 Special Issues

3.1 Wh-Phrases

Wh-phrases are not specially marked. It is of course important to distinguish wh-phrases from non-wh-phrases, but here the distinction is considered featural rather than categorial, hence outside the scope of the present evaluation.

[$_{\text{NX}}$ who]
[$_{\text{NX}}$ what]
[$_{\text{NX}}$ the man] [$_{\text{NX}}$ that] [$_{\text{NX}}$ you] [$_{\text{VX}}$ saw]
but: [$_{\text{NX}}$ I] [$_{\text{VX}}$ think] that [$_{\text{NX}}$ you] [$_{\text{VX}}$ saw]
in [$_{\text{NX}}$ whose house]
[$_{\text{AX}}$ how big]
[$_{\text{RX}}$ how quickly]
[$_{\text{VX}}$ know] when [$_{\text{INF}}$ to quit] *(*when *is not multi-word)*

3.2 Complementizers

Complementizers are not marked at all. Participles that seem lexically part of a complex complementizer are nonetheless treated as participles:

[$_{\text{VGX}}$ considering] that
[$_{\text{VNX}}$ provided] that

3.3 Punctuation

Punctuation is only included in chunks when it is "trapped" between other material:

[NX the " New Deal] " [NX that] [NX Roosevelt]

3.4 Coordination

Chunks do not contain coordinators unless they are "trapped" between other pieces or embedded in a premodifier. For example:

[NX we] [VX lack] [NX the ways] and [NX means] [INF to do] [NX that]
[NX the Ways and Means Committee]
[NX many] of [NX Georgia] 's [NX registration and election laws]

To amplify, we assume coordination to be a very different process from normal constituency. In particular, WE DO NOT TAKE SEMANTICS AS A GUIDE TO SYNTACTIC STRUCTURE WHERE COORDINATION IS INVOLVED. This assumption is admittedly nonstandard, but permits us to avoid difficult semantic judgments. For example, though *women* is semantically coordinated with *men* in (one reading of) *the old men and women*, we take the syntactic structure to be

[NX the old men] and [NX women]

When interpreting such a structure, we must recognize that *women* is in the scope of both *the* and *old*, but we take that to be semantic and not syntactic scope, or at least not syntactic scope at the chunk level.

- In clear cases of non-constituent coordination (e.g., right-node raising) the first piece is treated as involving an ellipsis:

  [NX the old] but not [NX the new styles]
  [NX the old but not new styles]
  [VX might have] and [VX certainly ought] [INF to have won]

- Precoordinators (e.g., *both* in *both...and*) are treated as coordinators. They are included in chunks only when trapped:

  both [NX John] and [NX Mary] [VX left]
  [VX will both try] and [VX succeed]
  in both [NX John] 's and [NX Mary] 's [NX opinion]
  not [RX so much] [AX difficult] as [AX tedious]
  [NX it] [VX was not so much difficult] as [AX tedious]

  But be careful to distinguish precoordinators from predeterminers: *[NX both (the) men] [VX left].*

- List commas are treated as coordinators.

  [NX a man] , [NX woman] , and [NX child]

- But: written-out numbers are not considered to involve coordination:

  [NX six million, four thousand and twenty-three (men)]

- Words connected with hyphens or slashes are considered to be compound words, not coordinated: *[NX singer/song-writer] [NX Billy Joel].*

3.5 Parentheticals

Full parentheticals, set off by dashes or parentheses, break up chunks:

of [NX local] – and I'm sure he realizes this – [NX governments]

3.6 Prepositions

Prepositional phrases are not marked. Prepositions generally stand alone, unless they are "trapped" in a larger chunk.

- Participles, nouns, etc., that appear to function as prepositions are nonetheless treated as participles/nouns/....

  in [NX spite] of [NX it]
  [VGX according] to [NX the latest figures]
  [VGX considering] the [NX trouble] [NX we] [VX 've been] to

- The same goes for *worth* and *missing: [AX worth] [NX the effort], [VGX missing] [NX a tooth].*

- *As* is ambiguous between a degree word and a preposition. The degree word is included in AX/VGX/VNX, but the preposition is not—unless it is trapped (cf. Measure Phrases, below).

  [AX as big] as [NX a house]
  [NX the values] as [VNX determined] by [NX Wiener] 's [NX method]
  [AX as much as a meter longer]

3.7 Partitives and Measure Phrases

Measure phrases are considered to belong to a level below the chunk level. Hence they do not necessarily constitute separate NX's. Note that they can modify nouns, adjectives, adverbs, and prepositions:

[$_{NX}$ a dozen men]
[$_{NX}$ twice the effort]
[$_{NX}$ two parts gin]
[$_{NX}$ a half a dozen men]
[$_{AX}$ two minutes early]
[$_{RX}$ two minutes earlier/ago]
[$_{NX}$ a mile] down [$_{NX}$ the road]
[$_{NX}$ the ship] [$_{VX}$ lies] [$_{RX}$ a mile down]
[$_{NX}$ three times more power]

- *As, than,* and coordinators can be embedded in these constructions:

  [$_{AX}$ six and a half times as expensive]
  [$_{NX}$ two times as much effort]
  [$_{NX}$ half again as many men]
  [$_{NX}$ six and a half times as expensive a project]
  [$_{NX}$ more than two crows]
  [$_{VX}$ increased] by [$_{NX}$ more than a third]
  [$_{NX}$ it] [$_{VX}$ was at least as effective]

- Though the meaning is different, we treat the *per/a* construction analogously:

  [$_{NX}$ 60 miles a second]
  [$_{NX}$ 60 miles per second]

- Measure phrases are taken to include adjectival and "bare" partitives:

  [$_{NX}$ how big a house]
  [$_{NX}$ what gauge wire]
  [$_{NX}$ how big a diameter circle]

- An *of*-PP is embedded only if it is trapped:

  [$_{AX}$ how big] of [$_{NX}$ a house]
  [$_{NX}$ how small of a gauge wire]
  [$_{NX}$ a quarter] of [$_{NX}$ a mile]
  [$_{AX}$ a quarter of a mile long]
  [$_{NX}$ most] of [$_{NX}$ a mile]
  [$_{NX}$ a quart] of [$_{NX}$ beer]
  [$_{RX}$ a quart of beer later]

3.8 Where NX Ends

Compound nouns constitute a single NX, but appositives constitute multiple NX's.

- Items with numbers or other designators are considered compounds:

  [$_{\text{NX}}$ Room 23B]
  [$_{\text{NX}}$ paragraph 16]
  [$_{\text{NX}}$ Ford Model T]
  [$_{\text{NX}}$ uranium 235]

- Both restrictive and non-restrictive appositives constitute multiple NX's:

  [$_{\text{NX}}$ former fire chief] [$_{\text{NX}}$ Marvin Dirtwater]
  [$_{\text{NX}}$ Marvin Dirtwater] , [$_{\text{NX}}$ the fire chief] ,
  [$_{\text{NX}}$ the " abortion pill] " [$_{\text{NX}}$ RU-486]

  Dates, addresses, titles, citations, mathematical and chemical formulae, and the like are considered compounds:

  [$_{\text{NX}}$ July 22, 1989]
  [$_{\text{NX}}$ the fourth of November, 1989]
  [$_{\text{NX}}$ Alabama, Georgia]
  [$_{\text{NX}}$ 12 Maple St., Alabama, GA]
  [$_{\text{NX}}$ Henry the Eighth]
  [$_{\text{NX}}$ Prof. John J. Jarvis, III]
  [$_{\text{NX}}$ Marcus (1980)]
  [$_{\text{NX}}$ a medium] [$_{\text{VGX}}$ containing]
  [$_{\text{NX}}$ 3.05×10^{-2} μM L-[*methyl*-^{3}H]-methione]

4 A Sample Text

Here is the beginning of sample A1 from the Brown Corpus, with chunks marked:

> [$_{\text{NX}}$ The Fulton County Grand Jury] [$_{\text{VX}}$ said] [$_{\text{NX}}$ Friday] [$_{\text{NX}}$ an investigation] of [$_{\text{NX}}$ Atlanta] 's [$_{\text{NX}}$ recent primary election] [$_{\text{VX}}$ produced] " [$_{\text{NX}}$ no evidence] " that [$_{\text{NX}}$ any irregularities] [$_{\text{VX}}$ took] [$_{\text{NX}}$ place].
>
> [$_{\text{NX}}$ The jury] [$_{\text{VX}}$ further said] in [$_{\text{NX}}$ term-end presentments] that [$_{\text{NX}}$ the City Executive Committee] , [$_{\text{NX}}$ which] [$_{\text{VX}}$ had] [$_{\text{NX}}$ overall charge] of [$_{\text{NX}}$ the election] , [$_{\text{VX}}$ deserves] [$_{\text{NX}}$ the praise] and [$_{\text{NX}}$ thanks] of [$_{\text{NX}}$ the City] of [$_{\text{NX}}$ Atlanta] for [$_{\text{NX}}$ the manner] in [$_{\text{NX}}$ which] [$_{\text{NX}}$ the election] [$_{\text{VX}}$ was conducted] .
>
> [$_{\text{NX}}$ The September-October term jury] [$_{\text{VX}}$ had been charged] by [$_{\text{NX}}$ Fulton Superior Court Judge] [$_{\text{NX}}$ Durwood Pye] [$_{\text{VX}}$ to investigate] [$_{\text{NX}}$ reports] of [$_{\text{NX}}$ possible irregularities] in [$_{\text{NX}}$ the hard-fought

primary] [NX which] [VX was won] by [NX Mayor-nominate] [NX Ivan Allen, Jr] .

Only [NX a relative handful] of [NX such reports] [VX was received] , [NX the jury] [VX said] , [VGX considering] [NX the widespread interest] in [NX the election] , [NX the number] of [NX voters] and [NX the size] of [NX this city] .

[NX The jury] [VX said] [NX it] [VX did find] that [NX many] of [NX Georgia] 's [NX registration and election laws] [VX are outmoded] or [AX inadequate] and [AX often ambiguous] .

[NX It] [VX recommended] that [NX Fulton legislators] [VX act] [INF to have] [NX these laws] [VNX studied] and [VNX revised] to [NX the end] of [VGX modernizing] and [VGX improving] [NX them] .

[NX The grand jury] [VX commented] on [NX a number] of [NX other topics] , among [NX them] [NX the Atlanta and Fulton County purchasing departments] [NX which] [NX it] [VX said] [VX are well operated] and [VX follow] [NX generally accepted practices] [NX which] [VX inure] to [NX the best interest] of [NX both governments] .