Content of Linguistic Annotation: Standards and Practices (CLASP)
Research Activities and Findings

Adam Meyers (editor), Nancy Ide, Charles Fillmore, Chris Cieri, Aravind Joshi, Martha Palmer, Nicoletta Calzolari, Nancy Ide, Rashmi Prasad, James Pustejovsky, Janyce Wiebe, Collin Baker, Bran Boguraev, Catherine Macleod, Critina Mota, Nianwen Xue, Dan Flickinger, Jan Hajic, Owen Rambow, Zdenka Uresova

2010

Introduction

25 members of the computational linguistics research community participated in a meeting at New York University on November 7, 2009 to address several difficult questions about the standardization of linguistic content in corpus annotation, where we define the term *standardization* to include all efforts to improve compatibility or interoperability between annotation content, including not only the creation of universal guidelines for particular types of annotation, but also any other type of *harmonization* efforts (e.g., mapping procedures that make two annotation schemes similar, projects involving correcting one type of annotation output based on another, etc.). It should be clear that *standardization* is a process rather than an end, in itself. Thus, this discussion aimed to establish recommended practices to further the cause of standardization, rather than a particular set of standards that should be adopted.

This workshop focused on the content of the annotation, rather than its physical format (xml, encoding issues, etc.) The scope of this workshop was complementary to efforts such as ISO's Linguistic Annotation Framework (LAF) [5], which focus on such issues.

This report summarizes the questions posed, the theoretical and practical considerations to be taken into account, the discussion that took place at the meeting, as well as some online discussions documented at:
`cims.nyu.edu/˜meyers/SIGANN-wiki/wiki/index.php/CLASP_Questions`
This is a portion of the CLASP website used for pre-workshop discussions of several questions relevant to the standardization process. Finally, this report summarizes both areas where the meeting participants reached consensus and areas where they did not.

1 The Pros and Cons of Standardization of Linguistic Content

Standardization of linguistic content aims to improve the interoperability of annotation created under different schemes, thus making it easier for single systems to use multiple types of annotation simultaneously. Some examples include:

1. Machine learning systems can more easily combine elements of annotation. For example, an ACE (`http://www.itl.nist.gov/iad/mig/tests/ace/`) system could use information such as whether an ARG0 of an *attack* verb tends to be an ARG1 of a *prosecute* verb (assuming PropBank's representation of predicate/argument structure [10]). This information is not detectable, however, unless the coreference and semantic role labeling (SRL) annotation are sufficiently interoperable–they must share some of the same basic units or it may be difficult to determine the coreference properties of any SRL arguments. Consider the following sequence of two sentences:

   Ms. Mary Smith assaulted the linguist. She was indicted later that day.

   Suppose that the coreference system detects that *Mary Smith* and *She* are coreferential and the SRL system detects that *Ms. Mary Smith* is an ARG0 of *assaulted* and *She* is the ARG1 of *indicted*. Unless an ACE system can reconcile that *Mary Smith* and *Ms. Mary Smith* are the same in some relevant way, this instance of a correlation between the ARG0 of *assault* and the ARG1 of *indict* cannot be used by this ACE system. This case could, in principle, be handled by a simple rule, e.g., the ACE system could assume that titles like *Ms.* are optional parts of names. However, there are many cases where name detection could vary between components and recognizing that a name is the same across these components can be complicated by many other factors such as the inclusion or exclusion of relative clauses, appositive elements and other modifiers, or names that are nested within other names.[1] Establishment of one system for identifying basic units that are used by both the NE classifier and the SRL system can eliminate this problem.

2. It is easier to merge annotation of the same variety if the annotation follows standards, e.g., if one wanted to train a part of speech tagger on a combination of text tagged with the Penn Treebank tagset and text tagged with the CLAWS tagset.

3. Merging several different annotation schemes into a single structure is easier when there are shared standards among the schema. This includes both: (a) systems like CONLL 2008/2009 and GLARF [11, 3, 8] that merge input annotation *aggressively* to force it to be compatible with a set of theoretical assumptions; and (b) systems like Ontonotes and MASC [4, 6], that merge passively, showing how the annotation lines up without changing it.

On the other hand, aggressively enforced standardization may hurt annotation research because versions of some theoretical analyses can simply not be made compatible with particular standards. Creators of the standards cannot always predict what is needed in the future, so this is unavoidable. However, good standardization practices can take these concerns into account. Under one reasonable approach, researchers would assume previous (standard) analyses unless they have a reason

[1] For example, systems vary with respect to whether *New York* is tagged as a name when it occurs as part of the larger name such as the *New York Yankees* or *the New York Museum of Modern Art*.

not to, i.e., researchers have the burden of proof for justifying new ways to analyze or represent phenomena. Documented standards would then be periodically updated to reflect such cases. In principle, following this philosophy would slow changes in linguistic analysis, but not prevent such changes when they were clearly necessary to describe some phonemenon. Still, it is unclear if members of the annotation community would ever willingly adopt such guidelines, or, indeed, whether any artificial incentives or disincentives should be implemented to influence annotators to do so. Work is being conducted all over the world in a wide variety of frameworks and under a wide variety of theoretical assumptions. If coerced, standardization can be seen as an imposition.

The CLASP workshop grew out of the Unified Linguistic Annotation (ULA) project (CNS-0551615 CRI-Towards a Comprehensive Linguistic Annotation of Language). Researchers from annotation projects including the Penn Treebank, PropBank, NomBank, TimeML, the Penn Discourse Treebank and Pittsburgh Opinion Annotation worked together to merge their annotation schemes together into a single representation and chose shared corpora for the purpose of annotation with multiple annotation schemes. In particular, the ULA work helped frame the harmonization problem along the lines described above.

2 Previous and Current Standardization Efforts

Several projects over the last 20 years or so have addressed issues of standardization for annotation content categories, including EAGLES/ISLE, which developed standards for content categories for morphosyntax, syntax, text typologies, subcategorization etc. All EAGLES/ISLE reports are available from http://www.ilc.cnr.it/EAGLES/isle/.

A list of some standardization efforts is on the CyberLing Wiki.[2] CyberLing is itself a standardization effort for data content categories (and other aspects of annotation) undertaken primarily within the linguistics community.

In particular, there are several ISO TC37 SC4 (Language Resource Management) Working Groups that have developed standards for: morphosyntax (ISO MAF), lexicons, syntactic annotation and time and events. There are also current ISO working groups to create standards for named entities, word segmentation (primarily Asian languages), spatial relations, among others.[3] Finally, there is a data category (www.isocat.org/) registry containing many linguistic categories and their definitions. Looking towards the future, this registry could be an important vehicle for standardization–the process of cataloging definitions as they are developed may encourage annotation researchers to make new definitions compatible with previous ones.

3 Administering Standards

The meeting broke down into four working groups, two of which focused on how standards should be carried out and the other two focused on how two specific phenomena might be standardized.

[2] cyberling.elanguage.net/page/Existing+Standards+and+Technologies

[3] See www.tc37sc4.org/new_doc/ISO_TC_37-4_N225_CD_MAF.pdf, www.lexicalmarkupframework.org/, www.tc37sc4.org/new_doc/ISO_TC37_SC4_N285_MetaModelSynAF.pdf and www.tc37sc4.org/new_doc/new_doc/iso_tc37_sc4_n269_ver10_wg2_24617-1_semaf-time_utf8.pdf

The Policy Working group and the Scope Working group discussed: what could (and should be done) to encourage annotation to take root; and what kind of phenomena should be standardized. These groups found that: (1) market forces, awareness-raising workshops, shared tasks, peer review and documentation requirements were the forces that could most effectively be used to foster the standardization process; and (2) any type of annotation that reaches maturity within a community is ready for standardization.

According to the Policy Working Group, market forces would cause systems that integrate smoothly to be preferred to ones that do not. For example, when annotation systems can be combined as part of a single shared task, each of those systems will be highly favored. These same forces will also cause people to produce adapters to map between systems to help ensure this sort of compatibility.

In the form of peer review and workshops, annotation researchers should be encouraged to: (1) create their own adapters to map their annotation to other frameworks; (2) provide detailed documentation; and (3) register annotation guidelines with the ISO registry.

Workshops for dealing with merging/compatibility issues are already popular and will continue. Furthermore, dissemination of horror stories about what happens when compatibility issues are ignored should help foster a desire for intercompatibility and the cooperation required to facilitate it.

According to the Scope working group, standardization should take the form of disseminating all annotation findings so that future annotation could be designed responsibly (without "reinventing the wheel"). Central repositories of documentation such as the ISOCAT registry should be a key part of the process.

Both of these working groups also emphasized standardization of the physical components of annotation, as per previous annotation efforts.

4 Two Candidates for Standardization

Large projects that incorporate many different types of automatic (or manual) annotation could be improved by the standardization of common components of the different systems. In particular, annotations that assume the same units (sentences, phrases, tokens, etc.) are easier to combine than those that don't. For this reason, we identified *tokenization* and what we call *anchor* selection as two areas of standardization worth exploring in detail, although other types of unit identification (sentences, text blocks, documents) may also benefit from standardization for the same reason. While the view that unit standardization is of particular importance ultimately turned out to be a minority position at the workshop, this idea is, nevertheless, the basis for discussing these particular aspects of linguistic structure at this meeting.

The consensus at the meeting seemed to suggest that the evolving standard practices would include several voluntary measures such as registering content categories with the ISOCAT registry. Thus we intend to add many of the considerations for anchor and token identification as recommendations in the ISOCAT registry for tokenization, dependencies and phrase structure.

By necessity, the guidelines proposed in this section (and in Section 6, the appendix at the end of this document) are English-specific. While these guidelines could, in principle, influence similar efforts for other languages, we decided that it would be a mistake to propose anglocentric guidelines for the world's languages.

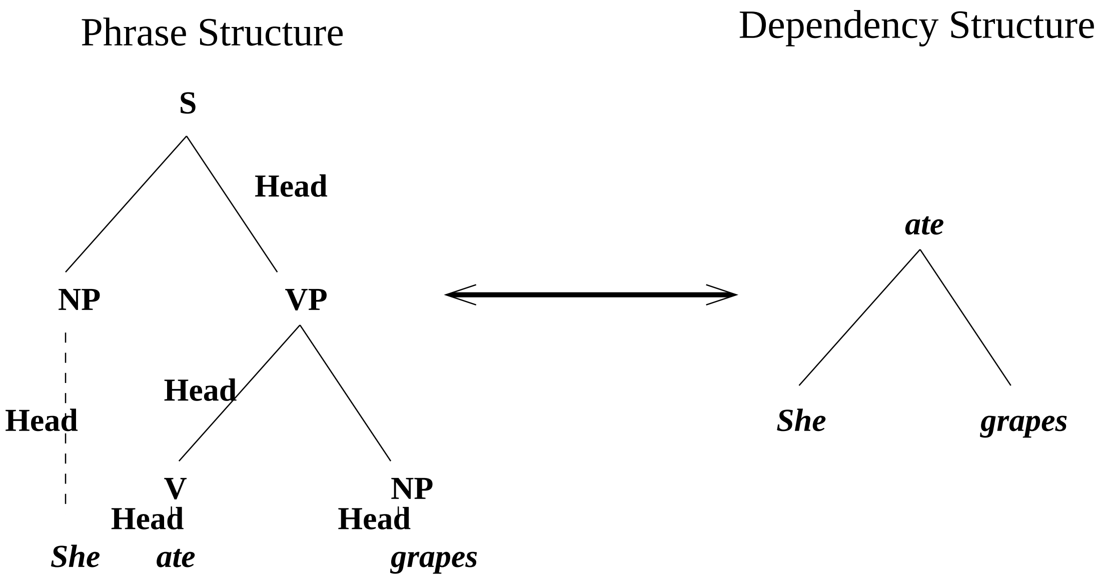

Figure 1: Mapping between Analyses of *She ate grapes*

4.1 Anchors

4.1.1 Defining the Anchor Problem

Most of the syntactic frameworks used in computational linguistics either break down sentences into phrases (a phrase structure approach) or identify which words "depend" on which other words (a dependency approach) or uses some combination of phrase structure and dependencies. It is widely recognized that these two types of representations have the following minimal differences:

1. A dependency analysis requires the selection of special words, typically called *heads* that dominate other items. Given a phrase structure analysis that require that heads are marked for every phrase, it is always possible to derive a dependency analysis. One need simply to promote the head of the phrase from its leaf position to the root of the highest phrase of which it is the head, in the process this highest phrase would be flattened, so that all dependents of the head would now be siblings in the new tree. Figure 1 shows this process graphically. Headless phrase structures cannot be translated into dependencies unless (provisional) heads are first selected by some criteria, e.g., those discussed in this section.

2. A phrase structure analysis distinguishes levels of embedding in a way that dependency analyses typically do not. For example, the concept of a verb phrase below the level of the sentence is not easily represented in a dependency analysis. Thus, dependency analyses tend to be equivalent to very flat phrase structure (S → NP V NP rather than S → NP VP) analyses. Of course nothing prevents variations of dependency analyses that would seek to represent this detail.[4]

[4] For example, the dependency grammar for Japanese assumed in the Kyoto corpus [7] assumes dependencies between small phrases called bunsetsus, rather than between words.

For purposes of using multiple resources, identifying that a particular unit is the same is crucial. Many researchers have found that standardizing relationships between heads (anchors) across different types of annotation is easier than standardizing relationships between phrases. For example, while various automatic analyses may assume different modifier attachments for noun phrases (NPs), NPs sharing the same head are probably the same. In other words, attempting to standardize the notion of "head" may be easier than attempting to standardize the notion of "phrase". Furthermore, the notion of "head" is shared by dependency analyses and the subset of phrase structure analyses mentioned in (1) above.[5]

Unfortunately, there are many types of phrases which don't really have heads, i.e., there is no single word that simultaneously determines: (1) the phrasal category of the phrase (NP, VP, etc.); (2) semantic features of the phrase, e.g., +/- human; (3) agreement features of the phrase (number, person); (4) selects the other items in the phrase; and (5) acts as the glue that links the other elements together. Such phrases (conjoined phrases, names, time expressions, and many others) either act like indivisible units (multi-word leaves), divide up characteristics 1–5, or otherwise do not conform to the only-one-head constraint.

Our standardization effort at the meeting involved the identification of a single item or set of items separated by white space which can standardly be used to represent these items. We called these representative item the *anchor* of the phrase, of the construction, (or of the dependency). We would argue that even if future standardization efforts reach different conclusions about the identification of the anchor, all such future efforts should be able to handle the list of difficult constructions that we describe here and in the appendix (Section 6) and more. Indeed, we would hope that our efforts would provide a jumping off point for future work in this area.

As noted above, there are several roles that the "head" of a phrase are suppose to play and one of the problems is that different roles are sometimes played by different constituents of a phrase. For example, the auxiliary verb *be* arguably, selects verbs as arguments, e.g., requiring that the following verb has either passive or progressive morphology (4 and 5 above), and auxiliary *be* also provides agreement properties of the phrase, e.g., *is* selects a third person singular subject. On the other hand semantic selection is based on the main verb, e.g., *John is eating.* is well-formed, but **The idea is eating.* is ill-formed, due to the semantic selection restrictions between *eat* and the subject. As we will discuss, any adequate account of anchors must: allow multiple types of anchors (akin to multiple levels found in many of linguistic theories); must provide a way for some anchors to inherit properties of one or more of their arguments; or must provide some other way of handling this mismatch (e.g., the "movement" analyses popular among theories based on the work of Zellig Harris, Noam Chomsky and others.)

There are some linguistic phenomena for which linguistic characterizations of phrases/heads/anchors are arguably irrelevant. These include topic analyses in terms of theme/rheme; representations of false starts, whispering, other speech phenomena, etc. Therefore, identifying phrases/anchors is only a partial solution to lining up different types of data. It is, however, extremely useful for a wide variety of phenomena. This is why it is worthwhile to work towards standardization of anchors.[6] Furthermore, even if a phrase lacks a theoretically-justifiable anchor, the phrase must still be represented somehow in whatever representation an NLP system is using. Otherwise, that

[5] Segmentation issues may complicate standardization of head selection, particularly for languages like Chinese, where segmentation decisions stir more controversy than they do in English. When segmentation is an issue, it may be easier to reach consensus on some unit larger (or smaller) than a word (a character, a phrase or a chunk).

[6] Some of these same issues were discussed in relation to the theory of syntax in [2, 9].

system will be unable to process that phrase. In an anchor-based system, such phrases must still be assigned anchors in a consistent manner.

4.1.2 Considerations for Choosing an Anchor

There are a number of things one should try to take into account when choosing an anchor for a phrase. The following two factors describe the function that an anchor ideally plays with respect to the other words that depend on it.

A. The anchor represents the phrase in the larger analyses (or dependency tree rooted at the anchor). Without identifying an anchor, a dependency graph representing the sentence would be a forest, rather than a rooted graph (or tree).

B. An anchor either: (i) represents the phrase's lexical properties, e.g., for purposes of selection (typically modeled as statistically based cooccurrence), or (ii) provides a predictable link to other lexical items that provide the phrase's lexical properties. For example, a really good definition of head (for this purpose) would capture that the phrase *eat and drink* has the lexical properties of the word *eat*, as well as those of the word *drink*.

The above considerations may not be sufficient to unambiguously choose a single word as an anchor. However, there are 2 common ways that many researchers choose an anchor in the absence of clear evidence:

C. Consistently choose either the first or last item. For example, conventions exist that would choose either *John* or *Smith* as the anchor of the phrase *John Smith*. This option is often chosen in order to make anchor selection easy to implement, consistent and predictable.

D. Create a multi-word anchor consisting of either all the words in the phrase or some privileged subset of those words. For various constructions including: fixed expressions (*down to earth*), so-called phrasal verbs (*call up*), and names (*John Smith*), it is sometimes assumed that these phrases are essentially words that contain spaces and the anchor is formed by concatenating the words together (*down to earth, call up, John Smith*). For multi-word solutions to work, consideration must be given to whether or not to include marginal words like titles in *Mr. John Smith* and how to handle cases where parts of the (hypothetical) anchor may have other non-anchor words interposed, e.g., the verb *call* and the particle *up* can be divided by the direct object, as in *John called Mary up on the phone*.

In addition to the considerations above, the amount of overhead necessary to identify a consistent anchor is also of concern since NLP systems vary with respect to the amount of processing that they do to identify an anchor (chunking, parsing, etc.) This factor may make it especially difficult to harmonize the output of high-overhead and low-overhead systems if both assume some version of the anchor concept.

4.1.3 Summary of Findings

The Anchor working group report included the following guidelines for choosing an anchor:

- Choose words as anchors, not morphological items.
- Aim to produce a connected graph of anchors for each orthographically defined sentence, possibly ignoring parentheticals.
- Never use null items as anchors, except possibly when it is unavoidable, e.g., gapping constructions.
- Choice of anchor should attempt to account for: agreement phenomena, case assignment and predicate argument structure.
- Difficult constructions should be dealt with.

The discussions that occurred on the Wiki page over the period of a month or so before the meeting provided more detailed specifications including: (1) discussions of many of the problem cases, which the Working Group agreed should be dealt with by all adequate accounts; (2) a discussion of multi-level analyses in which different types of anchors (surface and logical) are posited; and (3) the handling of filler/gap constructions. We have provided a detailed description of these findings as an Appendix (Section 6) at the end of this report.

4.2 Tokenization

A token is the smallest linguistic unit for some level of linguistic description. Tokenization standards can be developed according to several different strategies, each of which raise raise slightly different questions about standardized tokenization.

- **String Preserving Strategy**: Tokenization must preserve substrings of the original text in the tokenization
- **String Mapping Strategy**: Tokenization is a function that maps substrings of the text onto tokens

Another consideration is how tokenization should be defined, or more specifically, should tokenization include string regularization (mapping alternative spellings, identifying immutable idioms, etc.). Consider the following derivational sequence:

1. *Big Blue's stock is going topsy turvy*
2. *Big + Blue + 's + stock + is + going + topsy + turvy*
3. *IBM + 's + stock + is + going + topsy turvy*
4. *IBM + 's stock + be + 3rd-sing-present + go + progressive + topsy turvy*

Let's suppose that: 1 is the original string; 2 and 3 are possible tokenizations; and 4 is an approximate next step (morphology). On one view, step 2 is a necessary intermediate step called tokenization which future analysis should be based on, e.g., 3 is derived from 2. On another view, 3 is really the level of tokenization and 2 is a useless and necessarily inconsistent level of representation. Of course there are intermediary positions which would include aspects of 2 and 3. Crucially, 2 through 4 can be viewed as offset annotation: pointers to 1 plus additional information, e.g., in 3, "IBM" is a label on annotation associated with the text span between position 0 to position 8 (*Big Blue*).

Issues in tokenization include the following issues in English, among others.

- Can a single character be part of multiple tokens? For example, is the period both an end of sentence marker and part of the token *etc.* in the following sentence?

  The arc contained: goats, chickens, mice, roaches, amoeba, etc.

- How should contracted forms be split apart? For example, does *can't* become: *can* + *n't*; or *can* + *'t*; or *can* + *not*.

- Do multiword units form single tokens or multiple tokens, e.g., is *Big Blue* one token or two?

The working group decided that the problem was a question of defining what is part of the level of tokenization and what is part of some other higher levels of representation (morphology, text regularization, coreference, etc.). They ended up recommending that tokenization be treated as being very closely tied to the orthography and that further levels were necessary to deal with the phenomena describe above. Specifically, they proposed the following constraints on tokenization:

1. No character can be part of two tokens

2. No character should be mapped to another in tokenization step

3. Each character is part of a token

4. Ordinary space is not a token for English, but

   - Location of spaces is significant
   - Stop-start info must be retained as features
   - Types of whitespace are significant (line breaks, tabs)

Furthermore, they maintained that the following phenomena should be part of higher levels of processing:

- normalization of contracted forms

- correction of misspelled words

- normalization of alternative spellings

- aliases for named entities

- identification of immutable multiword units
- morphological analysis (inflectional or derivational)

In addition, special purpose grammars would be required to handle problematic cases such as: hyphenation, numbers, URLs, chemical names, abbreviations like *etc.*, metadata for later processors.

This approach has the benefit that it includes easy to agree upon aspects of analysis.[7]

Should these guidelines be adopted, however, the higher level of analysis which includes normalization of contracted forms, etc., would still have some of the issues discussed above. Thus, additional standardization would be required.

5 Concluding Remarks

In summary, the results of the meeting include the following:

1. Content standardization can only be the result of market forces and peer review.
2. The scope of content standardization includes all types of linguistic analysis that are sufficiently mature to become widely used.
3. Good documentation, especially when included in repositories such as ISOCAT will help facilitate good content standardization practices by the community.
4. If too restrictive, the enforcement of standardization can inhibit research.
5. Standardization of tokenization should be limited to the very basic distinctions that most people can agree upon. Most of the controversial issues (regularization of contractions, morphology, etc.) should be relegated to higher levels of analysis.
6. Accounts of anchor selection should have the following features in common: (a) to the extent possible, anchors should account for predicate argument structure, agreement and case assignment; (b) anchor selection should provide a set of dependencies that connect all the words (not null items) in a sentence; and (c) difficult constructions should be handled in some way.

6 Appendix: Details about Anchors

6.1 The Problem Cases

There is substantial agreement about the identity of the heads of most kinds of phrases: most NPs, VPs, PP, ADJPs, etc.[8]. For this reason, we will focus on the problematic cases, where anchor selection is not obvious.

[7] The Tokenization Working Group suggests that for languages like Japanese, Chinese and Arabic, tokenization would be very simple, but not so informative. Word Level detection which would operate on the tokens to produce words, based on linguistic analysis and dictionary entries.

[8] This ignores some of the theory-laden proposals to choose unconventional heads of phrases, e.g., analyses of noun phrases where the determiner is assumed to be the head of all phrases [1]. These have been popular among

For each case, we will outline why the case may be problematic and outline the range of possible solutions. The CLASP participants voted not to formally recommend a standard way of choosing an anchor, but nevertheless agreed that any adequate account should attempt to handle all the problem cases. We hope these clarifications will help future annotators and other NLP researchers find consistent solutions to the anchor problem.

Problem Case #1: The Anchor of a simple clause

Given a sequence of verbal elements in a simple sentence (marked S, SQ or SINV in the Penn Treebank), what is the anchor? Typically, the anchor is assumed to be either: (a) the main verb; or (b) the first auxiliary element. To clarify, auxiliary elements in English include: infinitival *to*, modals (*can, could*, etc.) and helping verbs *have, be and do*. The main verb is chosen because: it is the source of selection restrictions (e.q., *leave* selects subjects that can move and are non-abstract. Thus *a car*, *a person*, but not *an idea* is well-formed as the subject of *leave*). On the other hand, the first modal or auxiliary is the item that carries subject agreement inflection and/or determines key features of the sentence like tense and modality which can be selected for by superordinate predicates. For example, *want* selects sentential complements that are infinitives beginning with the auxiliary element *to*. Also, each auxiliary, modal or *to* selects something about the following one, e.g., *to* and modals selects bare verbs, *be* selects -ing or passive forms of verbs, *have* selects past participles, etc. These properties are documented throughout the linguistics literature. Some examples follow.

1. *[John left]* [*left* is the only possible anchor]
2. *[John might have left]* [The anchor is *might* or *left*]
3. *They want [to leave]* [The anchor is *to* or *leave*]
4. *Mary doesn't likes potatoes, but [Sam does]* [The anchor is *does, likes* or a gap bound by *likes*]

Of the examples above, the most difficult example for the main verb analysis is the last because resolving a verbal gap is beyond the level of processing that most NLP groups will complete. However, in addition, to either always choosing the first verbal element or always choosing the main verb, there are various compromises that can be adopted, e.g., choose the sequence consisting of auxiliaries and the main verb.

The advantage of main verb analyses include: (1) they are compatible with verb group heuristics assumed in many NLP systems; and (2) they focus on the lexical item with the most semantic content – for systems learning/using selection restrictions, it is useful to acquire co-occurrence statistics of main verbs and subject nouns, but not auxiliaries and subjects. Under the other approach, additional mechanisms can be used to connect subjects to the main verb. Since auxiliary items are a small class of items, a list of such items is easy to compile and can be used to link subjects to main verbs for purposes of selection. For example, the syntactic theory underlying the

some syntactic theories including Chomsky's Principles and Parameters theory and Dick Hudson's Word Grammar (a version of Dependency Grammar).

Penn Treebank could use an empty category in the subject position of the main verb that is bound to the "surface" subject.

Choosing the first verbal element as the head conforms to the syntactic structure of the sentence assumed by most linguistic theories, treebanks and the corresponding parsers. From a theoretical standpoint, tests for constituencies that have been used at least since the 1950s favor this approach. Some other advantages include: (1) the tense/mood of the clause is determined by the first verbal element and this is what complement selection by clause type is based on, i.e., when the clause is subordinate, verbs, adjectives and nouns select infinitive complements, subjunctive complements or conditional modals, or tensed complements; (2) the binary structure resulting from assuming the first verbal element is the head transparently accounts for sequences of verbal elements, e.g., in *the book may have been taken*, the modal *may* selects the bare verb *have* which selects the past participle *been*; and (3) conjunction can be handled in a straight-forward way when an auxiliary is distributed between conjoined verbs, e.g., in *John may not admit to the crime or accept the punishment*, the distribution of *may not* over both *admit* and *accept* follows easily from an analysis in which the VP *admit to the crime or accept the punishment* is a complement of the modal *may*. In contrast, the verb group style analysis in which the modal is subordinate to both verbs could get quite messy (and for example, could require gaps to represent the modification of the second verb by the modal).

For annotation purposes, a two level approach has been suggested (similar to the Penn-Treebank-style empty category approach mentioned above), in which the first verbal element should be considered the "surface" head and the main verb would be considered the "semantic" head.

Problem Case #2: The Anchor of a clause introduced by a subordinator

For sentences introduced by a SUBORDINATOR – a complementizer, question word or relative pronoun, there are two possible anchors: 1) The anchor of the sentence (as discussed under Problem Case #1) or 2) the SUBORDINATOR. These phrases are marked in the English Penn Treebank as either SBAR or SBARQ. The bracketed phrases in the following sentences are instances of clauses introduced by the subordinators in bold:

1. *[***What** *did I say?]*

2. *I said [***that** *I like cheese]*.

There are two common choices:

A. Assuming that the anchor of the simple clause is anchor of the subordinated clause has the following advantages: (1) it deals consistently with all subordinate finite clauses, e.g., the (bracketed) complement clause in *I said [I like cheese]* would have the same head as the **that** clause in the example above; (2) a simple analysis of relative clauses and questions would be possible in which the subordinator would only be an argument of the clause (the gap filler), leaving the role of the anchor to the main predicate.

B. Assuming that the subordinator is the anchor of the subordinated clause allows one to differentiate different types of clauses for selectional purposes. This may be an advantage over choice A. For example, whether the clause is introduced by a WH element, *that* or *for* distinguishes the clause with respect to complementation by the verbs: *said, ask, wonder* or the nouns *question, decision, desire*. Even when the clause is not a complement (e.g., a relative clause or an adverbial modifier), the subordinator is a major clue in determining the type of clause and thus helps with any distributional analysis. A disadvantage of choosing the subordinator as the anchor is that the word *that* is optional in the subordinate clauses that it introduces. There are two possible ways of handling *that*-less constructions: (1) when *that* is missing, adopt the anchor of the simple clause as the anchor; (2) insert an invisible place-marker for *that* (the manual version of the Penn Treebank inserts such an invisible element). However, choosing a gap as an anchor is undesirable from a theoretical point of view. For an element to be an anchor, it must be present. Inserting a gap to make it present when it is not makes the claim about anchors unfalsifiable.[9][10] The anchor working group specifically recommended not using invisible elements as anchors, with verbal gapping as a possible exception.

The CONLL 2008 and 2009 shared task for English differentiated *that* complements from relative clauses. In the former case *that* was assumed to be the anchor, but in the latter case the main predicate was assumed to be the anchor because the predicate takes the relative pronoun as an argument.

Problem Case #3: Names

Although names are NPs (as defined by distributional criteria), names that do not follow the determiner+adjective+noun paradigm must be treated specially because they are different (on the inside) than NPs headed by common nouns. We are aware of at least two conventions: (1) standardly choose the first (or last) token in a name as its anchor; and (2) treat names as words containing spaces, so that the entire name is not subdivided further for purposes of processing. Each of these conventions have variants in which a carefully designated subset of the tokens in the name are considered for anchor-hood and the other constituents are assumed to be modifiers. Specifically, titles like *Dr., President, Ms., Mr.,* . . . and post-honorifics like *Ph.D., Jr., III,* . . . are often omitted from people names. Sometimes post-honorific-like endings of company names like *Inc., Corp., Ltd., S.P.A.,* . . . are also omitted. Thus, for *Dr. John Smith, Jr.* and *Acme Products, Inc.* only *John Smith* and *Acme Products*, are considered. For this example, the CONLL 2008 and 2009 shared tasks for English would assume that *Smith* and *Products* were the anchors (a version of choice 1 which excludes such modifiers) and some other projects would assume that *John Smith* and *Acme Products* were two words containing spaces. Under both approaches, *Dr.* and *Inc.* were assumed to be modifiers of the names.[11]

Names and time expressions which have normal phrase structure (of an NP, ADVP, etc.) can, in principle, be treated as normal phrases, e.g., in the phrase *Association for Computational Lin-*

[9] This is also a major problem with some views of headedness adopted in the linguistic literature. For example, choosing the determiner as the head of the NP is odd, in part, because one has to insert invisible determiners into bare plurals (*pickles*) and determiner-less mass noun phrases (*water*) in order to maintain that the determiner is the head of these phrases.

[10] Other gaps are different. They provide a way of modeling that one construction essentially paraphrases another. For example, the gap in *The book was read* __*by Mary* indicates the relation of that sentence to the semantically similar *Mary read the book*. The gap in these type of examples are representational devices and are not being used to make square pegs fit into round holes.

[11] A complete grammar of names may require that the first and last names anchor dependencies independently of each other. This is exemplified by the following analysis of the phrases denoting *Randolf Quirk*, a British Lord. Specifically, the title *Sir* is licensed by the first name as demonstrated by the well-formedness of *Sir Randolf Quirk* and *Sir Randolf*, but the ungrammaticality of **Sir Quirk*. In contrast, the title *Lord* is licensed by the last name, as evidenced by the well-formedness of *Lord Randolf Quirk* and *Lord Quirk*, but the ungrammaticality of *Lord Randolf*.

guistics, we could assume that *Association* is the head. In other words, *Computational* depends on it Linguistics, which depends on *for*, which depends on *Association* and that *the* also depends on Association. Thus, we would be recognizing the consistent pattern that these phrases share with “ordinary” common noun phrases. However, it is also possible to treat these just like other names (anchored by the final word or by a set of words) for the purpose of consistency across names. Under this view, the whole phrase would either be treated as a word with spaces (Anchor = *Association for Computational Linguistics*) or the last three words would each depend on the first (assuming the first-word as anchor approach).

Problem Case #3: Time Expressions and Other Patterns

It is well-known that dates, times of day, numeric expressions, addresses, and other similar expressions have special grammars which may not include one dominant anchor-like item – these grammars are often consciously constructed by human beings (e.g., URLs). Compiling a detailed list is outside the scope of the meeting (and this report). However, such a list would be advantageous to the standardization process.

As with names, the anchor of a patterned expression would typically be either the whole expression (words with spaces) or the first (or last) token. Furthermore, in many cases the patterned expression can be distinguished from its modifiers (which need not be included in the anchor).

Some example patterned expressions follow.

Dates can consist of some subset of years, seasons, months, days of the week, days of the month and special year classifiers like BCE. However the date can be modified by adverbs and PPs (in bold): **precisely** *January 31, 1955*, *January 31, 1955*, **at night**, etc.

Numeric Expressions consist of sequences of numbers *five hundred thirty* and are separable from modifiers like *approximately* and unit expressions like ($, %) which typically act like head nouns, e.g., *$ 5* has the same meaning as *5 dollars*. In the latter case, it is uncontroversial that *dollars* is modified by *5*, but, in the former case, only some frameworks (e.g., the Penn Treebank) treat the number as a modifier of the unit punctuation.[12] When a number expression is written out as a sequence of words, it may also include the conjunction *and* conveying the same meaning as the word *plus*, e.g., *one hundred and fifty*. It is possible that *and* should not be included among the set of anchor words.

Expressions representing times of day, spelled out as word sequences may actually follow a head modifier pattern (similar to common nouns). For example, perhaps the bold-face items should be treated as anchors in the following examples: *half past* **three**, *quarter to* **one**, *ten to* **seven**, **eight** *fifteen*, *two* **o’clock**. Any set of guidelines should address this issue either by stating their assumptions for the full set of cases, or alternatively, treating the whole time expression (but not modifiers like *approximately* or *tonight*) as unanalyzable wholes.

Other examples include addresses (*1313 Mockingbird Lane*), court cases (*Roe vs. Wade*), degrees latitude/longitude, among others. An adequate system of anchor identification would need to standardize its treatment of all such phrases.

Problem Case #4: Verbal Particles

There are two common assumptions about verb particles, e.g., *up* in *Mary looked up the address*.

[12] The Penn Treebank uses an empty category to regularize *$5* so that it has the same interpretation as *5 dollars*.

1. **Verb-only approach**: A verb particle (*up* is dependent on the main verb (*look*). The main verb is the anchor of the VP.

2. **Phrasal-verb approach**: A verb plus its modifying particle form a complex verb. This complex verb is the anchor of the VP.

Phrasal-verb approach helps transparently accounts for the fact that verb plus particle constructions typically have different inventory of senses and subcategorizations than the verb does by itself.[13] One difficulty with the Phrasal-verb approach is accounting for the frequent discontinuity between the verb and its particle, e.g., *Mary looked the address up in the telephone book*.

Thus the choice between these options depends on what one considers to be a prettier theory. The Verb-only approach requires that one essentially allow the syntax of the verb (the presence/absence of a particle) to influence the possible sense inventories, whereas the Phrasal-Verb approach requires that one accepts a discontinuous words.

Problem Case #5: Simple Idioms

At the risk of opening up a can of worms, we will assume that there is a subset of idiomatic expressions that can be handled very similarly to names with respect to anchors. Specifically, we will consider the case of idiomatic multi-word expressions that omit no internal modification such as *wishy washy*, *cure all*, *air conditioner*, *blow dry*, *sort of*, *tip top* and *with respect to*. At this time, we will specifically refrain from writing up some standardization suggestions for the more difficult to characterize varieties of idiomatic expressions. However, clearly these should be dealt with as well.[14]

There are two possible options for the “simple” idioms:

1. **Idiom-as-name approach:** Treat them the same way that names are treated, either as words without spaces or assuming that the first or last word is the anchor.

2. **Idiom-as-non-idiom approach:** Create a standard analysis based on the syntactic environment that most closely fits that construction. For example, *wishy* is treated as a prenominal modifier of *washy*; *blow* is some kind of special modifier of the verb *dry*; etc.

The Idiom-as-name approach has the advantage of being conceptually simple. However, the Idiom-as-non-idiom approach scales better to handle some of the more complex types of idioms (that we are not dealing with here). For example, analyzing *keep tabs* as a verb object construction makes it easier to account for the various ways the idiom can be modified and altered, e.g., *He kept careful tabs on Mary; Tabs were kept on Oscar;* etc. On the other hand, the Idiom-as-non-idiom approach also forces the grammar to commit to ad hoc analyses, e.g., *blow* can only be a premodifier of the verb *dry* and no other verbs.

Problem Case #6: Coordination

[13] The set of possible subcategorizations may be as extensive as the possible subcategorizations of main verbs (without particles).

[14] Note the ungrammaticality (with the intended meaning) of **wishy very washy, *cure completely all*, **blow quickly dry*, etc.

In coordinately conjoined structures, all complete analyses must account for the following factors:

1. Coordinate conjunctions (*and, but, or, nor*) serve to link conjuncts, head-like phrases that jointly determine the lexical properties of the phrase,e.g.,

   [S [NP John and Mary] [VP ate and slept]]

   This sentence contains two coordinate structures. In both cases the conjunction is an instance of the word and. The conjoined NP has two conjuncts: *John, Mary*; The conjoined VP has 2 conjuncts: *ate, slept*.

2. Coordinate conjunctions can sometimes be missing, but the entire phrases still acts like a coordinate structure and its conjuncts can be easily identified. Sometimes, in written text punctuation (hyphens, colons, semicolons, commas, dashes) can be assumed to play the role of the conjunction. For example, two sentences can be linked with a semicolon, e.g., *John ordered a baked potato; Mary ordered a salad; Fred ordered a grapefruit.*

Complete analyses of coordination must complicate the dependency picture in some way. The conjunction is anchor-like in that it is the item that links the other phrases together – it identifies the structure as a coordinated structure. However, the lexical properties derive from the conjuncts. There are several ways to represent these properties.

We will discuss two alternative analyses of conjunctions: (1) the *cc-as-anchor* analysis; and (2) the *cc-as-modifier* analysis.

Under the *cc-as-anchor* approach, the coordinate conjunction is treated as a *transparent* (or syntactic) head, inheriting lexical properties from the conjuncts (semantic content, subcategorization/selection, agreement, etc.). In this sense, the conjunction is treated like many other closed class items (auxiliary verbs, light verbs, complementizers, some prepositions, etc.) which serve the syntactic function of linking together content words. For example, in *John ate and drank, ate* and *drank* are transparently dependent on *and*, so *and* inherits the lexical features of both verbs. Thus while *John* is the subject of *and,* this instance of *and* "acts" lexically like a merger of *ate* and *drank*.

Under the *cc-as-modifier* approach, a coordinate conjunction is treated as a preposition-like element that links one head word to another (with a semantics similar to some senses of *with* in English). For example, in *John slept and Mary ate, slept* is the anchor of the sentence, with dependents *John* and *and*; *ate* depends on *and*; *Mary* depends on *ate*. Under this approach one of the conjuncts (in English, the first one) is treated as the head. In order to still recognize all the lexical properties of the phrase, a special routine must be adopted to recognize that conjunct dependents should be treated as parallel to the head (and part of all the same selectional relations). For example, in *John ate and drank*, the relation between *John* and *drank* is mediated by the relations between: *John* and *ate*, *ate* and *and, and* and *drank*. In some systems, these additional lexical properties are simply ignored.

There are also more complex issues. Consider, for example, the *respectively* construction: *John and Mary*, respectively, *ate* and *drank*. Accurate connections of predicates and arguments would require a further elaboration of the interpretations mechanisms described above. For the *cc-as-anchor* approach, a mechanism would be required to capture that *and* connects each verb separately to its predicate. In a similar way, the *cc-as-modifier* approach would have to provide individual paths to connect each underlying predicate/argument pair.

Some multi-word coordinate conjunctions, including *as well as* and *in addition to*, function internally as simple idioms (see Problem Case 5), while their external syntax is of the sort described here. Other multi-word coordinate conjunctions appear in two separate parts, the first part appearing before the first conjunct, and the second part between two conjuncts. Example include: *either* + *or*, *neither* + *nor*, *not only* + *but*, and others. As with the verb particle construction, there are two common points of view: (1) the second element is the anchor and the first element is a special type of modifier – this is analogous to treating the verb as the anchor and the particle as a modifier; and (2) these are instances of two part anchors.

Problem Case #7: Apposition and Affiliation constructions

An **apposition** construction consists of two complete noun phrases such that the first is modified by the second, in an *is-a* relation, e.g., the expression *John Smith, president of Acme Co.* has an interpretation something like: *John Smith, who is the President of Acme*.

We are aware of at least two analyses of these constructions: (1) the *head-modifier* analysis and (2) the *parallel-conjunction-like* analysis.

Under the *head-modifier* analysis, the first NP is assumed to be the anchor of the whole expression, e.g., *John Smith* is the anchor of *John Smith, president of Acme.* This, we believe, is the most commonly assumed analysis.

Under the *parallel-conjunction-like* analysis, neither item is assumed to be the head (e.g., some versions of the ACE guidelines). While this analysis provides has some advantages for dealing with anaphora, it is somewhat antithetical to choosing an anchor for the phrase, since it basically assumes that neither is the anchor.[15]

An **affiliation** construction consists of two adjacent named expressions, separated by punctuation, e.g., *Trenton, N.J.*. While superficially similar to the apposition construction, there are the following differences: (1) two names can not normally participate in apposition, e.g., *Clark Kent, Superman* does not work as an apposition construction. We assume that these two names are *affiliated* with each other e.g., Trenton, New Jersey or Adam Meyers, NYU mean something like: the Trenton that is associated with New Jersey and the Adam Meyers that is associated with NYU. We are only aware of one analysis of these in which the first name is assumed to be the anchor (Trenton and Adam Meyers) and the second, the modifier (*New Jersey* and *NYU*).

It can be difficult to distinguish title + name constructions from instances of apposition. *Mr. John Smith* undoubtedly consists of a title plus a name and *the actor, Charlton Heston*, is undoubtedly an apposition construction. However, the following cases seem to fall somewhere in between these two cases:

1. uncle Paul
2. my brother Sam
3. Professor of linguistics Noam Chomsky

[15] Suppose we assume: (1) that appositives are conjunction-like and (2) that something parallel to the cc-as-modifier approach is correct. Then, this analysis becomes essentially equivalent to the *head-modifier* approach.

4. the Deputy Undersecretary of Defense William G. Boykin

One test that might be helpful is to determine whether the first of the two NPs forms a complete independent NP. Under this approach *uncle Paul* would be considered a title plus name construction, whereas the others above would be considered instances of apposition. Still, there appear to be some shades of gray.

Problem Case #8: Range Expressions

Numbers, measure and time expressions combine to form composite phrases, the anchor of which, is not obvious. There are two varieties we will discuss: range expressions and *per* expressions.

Range expressions include the bracketed portions of the following phrases:

1. *[from 5 to 10] dollars*

2. *[5 to 10] dollars*

3. *[5 - 10] dollars*

4. *[January 1 to February 3]*

5. *[From January to October]*

There is not a lot of literature on how these should be analyzed. The Penn Treebank either analyzes these as to PPs combined:
(*[PP [PP from [NP January]] [PP to [NP October]]]*)
or as a flat structure (typically a QP), e.g.,
[NP [QP 5 to 10] dollars]
Neither of these structures indicate what the head of the structure should be. We will thus propose the following two analyses: the TO-AS-ANCHOR analysis and the CHOOSE-FIRST-EXTENT analysis, each modeled after one of the above analyses of conjunction. The TO-AS-ANCHOR analysis assumes the following:
assume that:

- The word *to* or the hyphen - anchor these constructions

- As with the coordination, these anchors are transparent and they inherit the lexical properties of the linked expressions (the numbers or dates, in these examples).

In contrast, the CHOOSE-FIRST-EXTENT analysis assumes that the first extent (possibly including *from*) is the anchor and the final *to* (or hyphen) plus NP is the modifier. The same considerations that go into the cc-as-modifier approach apply here as well.

Problem Case #9: Rate Expressions

Rate expressions either: (1) combine two units to form a rate, linking the two unit expressions with either the word *per*, or an indefinite determiner; or (2) combine one unit expression with a following quantificational expression like *each* or *apiece*. Some examples are:

1. *twenty dollars a day*
2. *twenty miles an hour*
3. *twenty cents apiece*
4. *forty dollars each*
5. *seventy miles per gallon*

In the syntax of the Penn Treebank, the structure is given as follows:

1. *[NP [NP twenty dollars] [NP-ADV a day]]*
2. *[NP [NP twenty miles] [NP-ADV an hour]]*
3. *[NP [NP twenty cents] [ADVP apiece]]*
4. *[NP [NP forty dollars] [NP-ADV each]]*
5. *[NP [NP seventy miles] [PP per gallon]]*

In each case, the dependency analysis that is the most compatible with the Penn Treebank would be one in which the first noun (*dollars,miles,cents*) is the anchor. and the second phrase (NP-ADV, ADVP or PP) is the modifier.[16] Given that a rate can be described mathematically as a division problem or a fraction, we will call this analysis the **numerator-as-head** analysis (implying that the denominator is the modifier). An alternative analysis would be to assume the determiner, *per*, quantifier (*each*), or adverb (*apiece*) is the anchor. These mostly closed class words license the construction and so could be legitimately thought of as transparent anchors, the same way that conjunctions can under the *conjunction-as-anchor* analysis.

As with many special constructions, choosing the correct anchor is not straight-forward. In particular, neither the numerator, nor the denominator is really semantically dominant for deriving the meaning of the phrase. Due to the ungrammaticality of internal modification, it is hard to find a definitive argument for breaking the phrases down into constituents at all, e.g., **miles per nasty hour* is a distinctly odd phrase. Finally, many rates are conventionalized either as abbreviations (mpg, mph, etc.) or other terms (hertz, flops), suggesting that these units are more than the some of their parts. On the other hand, choosing the function words *a*, *an*, *per*, *apiece*, etc. has some pit falls as well, since the resulting phrase does have the external distribution of a measure NP, just like *miles* or *hour*. This suggests that a **word-with-spaces** analysis is also plausible, a third option. Under this analysis, the string *miles per hour* is the anchor of *five miles per hour*.

Problem Case #10: Nonlocal dependencies

There are several cases in English where dependencies cross in the sense that given an X that premodifies some Y, there is some Z, a dependent of X that follows Y. Some examples follow in which X is underlined and Z is in brackets:

[16] The theoretical status of function tags like -ADV is not crystal clear. However, it is clear that an NP-ADV constituent is not intended as the head of the superordinate NP.

1. *an* **easy** *book [to read]*
2. **too** *big [to eat]*
3. *a* **bigger** *problem [than I'd like to admit]*

In each of these examples, there is a crossing dependency. In the first example, the adjective complement follows the noun. Thus within the NP *an easy book to read, easy* takes an infinitive complement, which follows the head noun. In the ADJP, *too big to eat*, the degree word *too* takes an infinitival complement that follows the adjective. Finally, in the NP *a bigger problem than I'd like to admit*, the comparative *bigger* takes a *than* complement, which follows the head noun.[17]

Most systems simply ignore these sorts of cases (and many system developers are unaware of them). They are difficult because the crossing dependencies complicate the relation between word order and dependency is complicated – systems must allow for these special cases (and most purely statistical models do not have a simple way of including them).

We assume that these kinds of examples (and others) may be more easily handled under an analysis that assumes multiple "levels" of anchor/argument dependencies. At the "surface" level, we would assume that all subconstituents are dependent on the head of the phrase. Thus in the first example, we would assume that *an, easy* and *to read,* all depend on *book* on the surface. However, at the "logical" level, we would assume that *to read* depends on *easy*. We can assume that the logical level would be some regularized level along the lines of the semantic dependencies of CONLL 2008, 2009, PropBank, NomBank, FrameNet, and other annotation frameworks. In this view, the surface representation would preserve the order that the elements appear in the string and the logical level would be closer to a representation of meaning (at the very least linking the items together based on what selects what).

Problem Case #11: Light/Transparent verbs

Often verbs seem to inherit the combinatorial properties of one of the post-verbal arguments. The most common cases include: support verbs, and verbs that take predicative arguments.

In the following examples, the subcategorized post-verbal phrases and selectional constraints on the subject are determined by the object of the support verbs *have, make, give* and *do*. In each case, the noun requires that the subject of the sentence be sentient and places subcategorizational restrictions on the other elements in the sentence, e.g., the *with* phrase in sentence 2. Furthermore, the alternation between the plural subject in sentence 1 and the subject plus the *with* phrase is limited to the combination of *have* plus a particular set of nouns {argument, fight, quarrel, . . .}.

1. *John and Mary had an argument*
2. *John had an argument with Mary*
3. *Mary made a derogatory statement to the Press about John*
4. *John gave Mary bad advice*

[17] Other kinds of degree, comparative and superlative complements include: *that clauses* as in *so hungry [that I cannot think]*; NPs as in *too big [a problem]*; and PPs as in *the tallest mouse [in the world]* and *too big [of a problem]*.

5. *Mary did John a favor*

A similar problem occurs with verbs like *be* and *seem*, verbs that are typically categorized as copulas or subject to subject raising verbs. In the following examples, the phrase following the verb selects properties of the subject and is mainly responsible for the meaning of the VP. In the examples 6 and 7, the subject must be a tangible object, capable of being tasted (due to *salty*) or located somewhere (due to the locative PP). Intangible subjects like *sincerity* would result in ill-formed sentences. Similarly, all complements are determined by the post-verbal predicate (even though most parsers and treebanks seem to attach them to the verb), as in examples 8 and 9.

6 *The food seems salty*

7 *Mary was on the train*

8 *John is angry at Mary*

9 *Mary is insistent that John stops all that nonsense*

In both of these cases, most analyses assume that the verb is the anchor of the verb phrase and/or the sentence and the other elements depend on it. As with the nonlocal selection case, perhaps a multi-level analysis may help here. In fact most frameworks that handle the selection and subcategorization properties of these kinds of constructions due assume multi-level analyses of some type.

6.2 The Surface/Logical Anchor Distinction

Several of the problem cases described above may become less problematic under a two "level" analysis which distinguishes as syntactic anchor from a logical one. The **surface** anchor serves the following functions: (1) it provides a convenient way to construct an analysis, regardless of whether a single element is truly head-like; and (2) it tends to include function words with little semantic content. The **logical** anchor has the following properties: (1) it semantically selects all its dependents; and (2) it must be a content word.

The following are some syntactic/logical anchor assignments, based on the analyses of some of the problem cases above. An item is in bold if the analysis in the text above marks it as the anchor of the construction.[18]

Problem Type	**Surface Anchor**	**Logical Anchor**	Example
Nonlocal Selection	Phrasal Head	Modifier	**too** *big [to eat]*
Coordination (cc-as-anchor)	conjunction	set of conjuncts	*John* **and** *Mary*
Coordination (cc-as-modifier)	first conjunct	set of conjuncts	**John** *and Mary*
Clause with Subordinator	subordinator	main verb	**that** *Mary is human*
Simple Clause	1st aux or verb	main verb	*Mary* **may** *leave*

[18] For the nonlocal selection example, we are only considering dependencies involving the clause *to eat*: the surface dependency between *big* and *to eat* and the logical dependency between *too* and *too eat*. The fact that *too* is dependent on *big* is not relevant.

It should be clear that the surface and logical anchors could be derived from the analyses above in these cases. For the most part, alternative surface analyses should result in the same choice of logical anchor even if they assume different surface anchors, at least for these cases.

For the light and transparent cases, one can make a similar sort of distinction, assuming that the support and copular verbs are the anchors at the surface level, but the nouns or predicative complements are anchors at the logical level. Under the contemplated analysis, *had, made, gave, did,* are surface anchors and *argument, statement, advice, favor, salty, angry, insistent* and *on*[19] are the logical anchors.

1. *John and Mary had an argument*
2. *John had an argument with Mary*
3. *Mary made a derogatory statement to the Press about John*
4. *John gave Mary bad advice*
5. *Mary did John a favor*
6. *The food seems salty.*
7. *Mary was on the train.*
8. *John is angry at Mary*
9. *Mary is insistent that John stops all that nonsense.*

Furthermore modifiers including subordinate clauses and some PPs, arguably have different logical and surface structures. Thus on the surface (e.g., the Penn Treebank), the *if* clause modifies the main clause in the sentence *I will leave [if you pay me]*, i.e., *if* depends on *will* or *leave*. However, under some logic-based analyses (the Penn Discourse Treebank), *if* takes both clauses as arguments (both *pay* and *will leave* depend on *if*).

In short, there are several cases outlined above in which a distinction between a surface and a logical anchor may make for a more natural analysis.

6.3 Elided or missing elements

Cases of missing, but implied elements are typically handled by so-called filler/gap analyses. For example, in *I want to leave*, the missing subject of the infinitive *leave* is assumed to be a gap that is filled by the NP *I*. By analogy to standard clauses like *I left*, so-called empty category analyses[20] would insert an invisible, anaphor-like entity called an empty category (*e*) immediately before the infinitive and assume that it is co-indexed (like a pronoun) with the subject of the dominant clause, i.e., I_i *want* e_i *to leave.* For some constructions, the most natural analysis would make a missing element the anchor of the phrase. Some examples follow:

[19] We will ignore the issue about whether the preposition or its object should be the logical argument of this sort of PP.

[20] Such analyses were made popular over the last fifty years or so by the work of Noam Chomsky and others.

1. *I want five red beans$_i$ and three green e_i?*
2. *Some [would take them to]$_i$ the top half of the dale and others e_i the bottom half.*
3. *She [could not say it]$_i$ in front of him nor he e_i in front of her.*
4. *You're [in no bigger of a hurry] than I am e_i to get this job finished.*
5. *Bears have become largely e_i and pandas entirely e_i noncarnivorous$_i$*

The working group responsible for anchors specifically stated that it was undesirable to posit null elements (empty categories) as anchors of phrases. On the other hand selecting any other element would also be problematic for these cases. This conflict makes these sorts of (luckily not so common) examples difficult for anchor selection.

Possible ways to handle these constructions include: (1) positing empty heads (contrary to the recommendations of this meeting); or (2) (equivalently) positing that the antecedent (in effect) is in some sense, the anchor that is implied by the position of the gap (for example at the logical level). Further complications include: (a) the gaps often represent type identity rather than token identity with their antecedent – the examples above refer to different instances of beans, *taking them to somewhere*, *saying it*; (b) not all the gap fillers form natural constituents, e.g., *would take them to* does not form a constituent under most grammars.[21]; (c) the word *nor* "absorbs" the negation portion of the meaning of the gap in (3). Thus being forced to "choose" an anchor is not ideal because these kinds of examples are good evidence that the type of analysis that requires anchors has holes in it.

6.4 WH Extraction

For relative clauses and WH questions, the relative or interrogative pronoun arguably serve two functions simultaneously: (1) they introduce a clause; and (2) they fill a gap in that clause. Some examples follows:

1. What$_i$ did you eat e$_i$?
2. The sandwich$_i$ that$_i$ you ate e$_i$?

In our discussion of problem case #2 described above, we discussed the possibility that the interrogative pronoun anchors the clause, even though the gap filled by the pronoun is clearly subordinate to the clause. The assumption that the pronoun is the surface anchor and the main verb is the logical anchor would be compatible with many accounts that make this sort of distinction.

A similar account is possible for the relative pronoun in a relative clause. However, the relative pronoun analysis must also include some sort of coreference relationship between the relative pronoun and the head noun modified by the relative clause. The specifics of the relative clause analysis are further complicated by examples where the relative pronoun is embedded in a larger phrase, such as:

3 the hill [[on the top of which] we first met]

[21] It does form a connected graph under most dependency analyses.

4 the book [[whose pages] are marked with Batman's insignia]

Such examples may lead one to reject the relative pronoun (or the dominating phrase) as the anchor, unless a more complex analysis is assumed.

6.5 Cases Where Anchors Cannot be Used

There are several speech and discourse-related phenomena which make it difficult, if not impossible, to maintain a pure anchor-based analysis. By pointing these out, we are not advocating rejecting anchor-based analyses. Rather, we are trying to clarify the limitations of making these extremely useful idealizations.

In speech analysis, it's often necessary to indicate stretches of speech that are whispered, in creaky voice, or in falsetto, say, and such labels don't necessarily stretch over what the syntactician would regard as constituents. In other words, this is a clear case when we need a linear record of an utterance form on one line, and on separate lines indications of the places where some prosodic or other features get turned on and off.

In a similar way, interruptions, parentheticals and other intrusions in speech are not always subject to the same syntactic constraints as "normal" syntactic units (or their dependency equivalents). The following examples are from the Contemporary Corpus of American English (Mark Davies):

1. We were firmly set in place, rooted, [as it were], to the spot
2. I don't think I'm at all [...how should I put it...] suggestible
3. it's not, [so to speak], in your purview
4. the - [how should I put it] - unusual division of responsibilities in this household

In these examples, the bracketed phrases behave as if they occur on a separate level from the rest of the sentence, e.g., *how should I put it* is not a normal part of a noun phrase occurring between a determiner and an adjective, but rather an independent sentence which represents comments of the main sentence's speaker.

References

[1] Steven Abney. *The English Noun Phrase in its Sentential Aspect*. PhD thesis, MIT, 1987.

[2] Greville Corbett, Norman M. Fraser, and Scott McGlashan. *Heads in Grammatical Theory*. Cambridge University Press, Cambridge, 1993.

[3] J. Hajič, M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, P. Straňák, M. Surdeanu, N. Xue, and Y. Zhang. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *CoNLL-2009*, Boulder, Colorado, USA, 2009.

[4] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. Ontonotes: The 90% solution. In *NAACL/HLT*, 2006.

[5] N. Ide and K. Suderman. Graf: A graph-based format for linguistic annotations. In *Proceedings of The Linguistic Annotation Workshop, ACL 2007*, pages 1–8, Prague, 2007.

[6] Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. Masc: the manually annotated sub-corpus of american english. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.

[7] S. Kurohashi and M. Nagao. Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of The 1st International Conference on Language Resources & Evaluation*, pages 719–724, 1998.

[8] A. Meyers, M. Kosaka, N. Xue, H. Ji, A. Sun, S. Liao, and W. Xu. Automatic Recognition of Logical Relations for English, Chinese and Japanese in the GLARF Framework. In *SEW-2009 at NAACL-HLT-2009*, 2009.

[9] Adam Meyers. The NP Analysis of NP. In *Papers from the 31st Regional Meeting of the Chicago Linguistic Society*, pages 329–342, 1995.

[10] M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

[11] M. Surdeanu, R. Johansson, A. Meyers, L. Márquez, and J. Nivre. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *Proceedings of the CoNLL-2008 Shared Task*, Manchester, GB, 2008.