Syntactic Annotation in the Columbia Arabic Treebank

Nizar Habash†, Reem Faraj ‡ and Ryan Roth†

†Center for Computational Learning Systems, Columbia University
{habash,ryanr}@ccls.columbia.edu
‡Middle East and Asian Languages and Cultures, Columbia University
{rf2273}@columbia.edu

Abstract

The Columbia Arabic Treebank (CATiB) is a database of syntactic analyses of Arabic sentences. CATiB contrasts with previous approaches to Arabic treebanking in its emphasis on faster production with some constraints on linguistic richness. Two basic ideas inspire the CATiB approach. First, CATiB avoids the annotation of redundant linguistic information that is determinable automatically from syntax and morphological analysis, e.g., nominal case. And secondly, CATiB uses linguistic representation and terminology inspired by the long tradition of Arabic syntactic studies. This makes it easier to train annotators and not be restricted to hire annotators who have degrees in linguistics. This paper describes CATiB's representation and compares it to other Arabic treebanking efforts.

1. Introduction

Collections of manually checked syntactic analyses of sentences, or treebanks, are an important resource for building statistical parses and evaluating parsers in general. Rich treebank annotations have also been used for a variety of applications such as tokenization, diacritization, part-of-speech (POS) tagging, morphological disambiguation, base phrase chunking, and semantic role labeling. Under time restrictions, the creation of a treebank faces a tradeoff between linguistic richness and treebank size. This is especially the case for morpho-syntactically complex languages such as Arabic or Czech. Linguistically rich representations provide many (all) linguistic features that may be useful for a variety of applications. This comes at the cost of slower annotation as a result of longer guidelines and more intense annotator training. As a result, the richer the annotation, the slower the annotation process and the smaller the size of the treebank. Consequently, there is less data to train tools.

In the case of Arabic, two important treebanking efforts exist: the Penn Arabic Treebank (PATB) (Maamouri et al., 2004; Maamouri et al., 2009) and the Prague Arabic Dependency Treebank (PADT) (Smrž and Hajič, 2006; Smrž et al., 2008). Both of these efforts employ complex and very rich linguistic representations that require a lot of human training. The amount of details specified in the representations is impressive. The PATB not only provides tokenization, complex POS tags, and syntactic structure; it also provides empty categories, diacritizations, lemma choices and some semantic tags. This information allows for important research in general NLP applications; however, much of this rich annotation is currently unused in Arabic parsing research (Kulick et al., 2006) since it is generally considered to be derivative of the output of parsing itself. For example, nominal case, which can be determined for gold syntactic analyses at high accuracy (Habash et al., 2007a), cannot be predicted well in a pre-parsing POS tagging step (Roth et al., 2008; Habash and Rambow, 2007).

In this paper, we present the Columbia Arabic Tree Bank (CATiB). CATiB contrasts with previous resources in putting an emphasis on faster production with some constraints on linguistic richness. Two ideas inspire the CATiB approach. First, CATiB avoids annotation of redundant linguistic information. For example, nominal case and state (definite, indefinite, construct) in Arabic are determined automatically from syntax and morphological analysis of the words and need not be annotated by humans. Of course, some information in CATiB is not easily recoverable, such as phrasal co-indexation and full lemma disambiguation. Second, CATiB uses a linguistic representation and terminology inspired the long tradition of Arabic syntactic studies. This makes it easier to train annotators, who need not have degrees in linguistics. CATiB uses an intuitive dependency representation and relational labels inspired by Arabic grammar such as tamyiz and idafa in addition to the well-recognized labels of subject, object and modifier. In this paper, we focus on describing the CATiB annotation guidelines for most common linguistic constructions. A full description is available in the CATiB manual (Habash et al., 2009). A discussion of CATiB annotation speed, inter-annotator agreement, parsing results and annotation process is presented in (Habash and Roth, 2009). In the next three sections, we describe CATiB guidelines for tokenization, POS tagging and syntactic annotation. Section 5. compares CATiB representation with both PATB and PADT's representations. Section 6. briefly describes CATiB's publicly released package.

2. Tokenization Guidelines

Words in CATiB are white-space and punctuation separated strings. Words are broken into tokens for annotation purposes in a manner inspired by the tokenization in PATB. Only the following clitics are separated from the word: +أ *Â*+[1] 'question particle', +و *w*+ 'and', +ف *f*+

[1] All Arabic transliterations are provided in the Habash-Soudi-Buckwalter transliteration scheme (Habash et al., 2007b). This scheme extends Buckwalter's transliteration scheme (Buckwalter, 2002) to increase its readability while maintaining the 1-to-1 correspondence with Arabic orthography as represented in standard encodings of Arabic, i.e., Unicode, CP-1256, etc. The following are the only differences from Buckwalter's scheme (which is indi-

'so/then', +ل *l+* (the preposition 'for/to', the conjunction 'so that' and the verbal modifier particle 'indeed/truly'), +ب *b+* 'by/with', +ك *k+* 'as/like', and pronominal (object/possessive) clitics such as +ه *+h* 'him/his' and +كم *+km* 'you/your'.[2] Since currently no choice of diacritization is made as part of CATiB, all diacritics are removed. For example, ولبيوتكم *walibuyuwtikum* 'and for your houses' is tokenized as و+ل+بيوت+كم *w+l+bywt+km* 'and+for+houses+your'. Since decliticization can sometimes lead to malformed word forms, we normalize the resulting tokens to their naturally uncliticized form. The following are some examples of this normalization:

- Alef-Lam: للكتاب *llktAb* 'for-the-book' becomes ل+الكتاب *l+AlktAb* (not ل+لكتاب *l+lktAb*)
- Ta-Marburta: مكتبتنا *mktbtnA* 'our library' becomes مكتبة+نا *mktbħ+nA* (not مكتبت+نا *mktbt+nA*)
- Alef-Maqsura: مستشفاهم *mstšfAhm* 'their hospital' becomes مستشفى+هم *mstšfý+hm* (not مستشفا+هم *mstšfA+hm*)
- Case-variant Hamza: بهاؤه/بهاءه/بهائه *bhAŵh/bhA'h/bhAŷh* 'his glory [nom./acc./gen.]' becomes بهاء+ه *bhA'+h* (not بهاؤ+ه/بهائ+ه *bhAŵ+h/bhAŷ+h*)

3. POS Tagging Guidelines

There are only six POS tags in CATiB. The tags are inspired by the traditional Arabic grammar classification of noun, verb and particle (اسم ، فعل وحرف). The simplicity of the POS tagset is intended to speed up human annotation yet maintain important distinctions. We discuss this further in Section 5.

- **VRB** is used for all verbs including the class of *incomplete verbs* (أفعال ناقصة), also known as *Kana and its sisters* (كان وأخواتها). Examples of incomplete verbs include كان *kAn* 'be', صار *SAr* 'become' and ليس *lys* 'be not'.
- **VRB-PASS** is used for passive-voice verbs.
- **NOM** is used for all nominals such as noun, adjective, adverb, active/passive particple, deverbal noun (مصدر), pronoun (personal, relative, demonstrative, interrogotive), numbers (including digits), and interjections. Preposition-like nouns/adverbs such as أمام *ÂmAm* 'in-*front*-of' and فوق *fwq* 'on-*top*-of' are considered NOMs. Similarly, quantifiers such as كل *kl* 'all' and بعض *bςD* 'some' are also considered NOMs.
- **PROP** is used for proper nouns. Given that Arabic does not explicitly mark proper nouns, we use the English capitalization guidelines for marking proper nouns as our guide. For example كريم *kariym* is NOM if it translates as 'generous', but PROP if it translates as 'Karim'. Similarly, all the words including بنك *bank* are PROP in بنك القاهرة عمان *bank AlqAhirah ςam∼An* 'Cairo Amman Bank'.
- **PRT** is used for all particles. This is a superset including the following different closed-classes:
  - Prepositions such as من *min* 'from', إلى *Ăilaý* 'to', عن *ςan* 'about' and على *ςalaý* 'on'.
  - Coordinating conjunctions such as +و *wa+* 'and', أو *Âaw* 'or' and ثمّ *θum∼a* 'then'
  - Subordinating conjunctions such as كي *kay* 'in order to', لكن *lákin* 'however' and أن *Âan* 'that'
  - Conditional conjunctions such as إذا *ĂiðA* 'if'
  - The class of *conjunctional verb-like particles* (حروف مشبهة بالفعل), also known as *Inna and its sisters* إنّ وأخواتها, which includes among others, إنّ *Ăin∼a* 'that/indeed', أنّ *Âan∼a* 'that' and لكنّ *lákin∼a* 'however'
  - The attention particle (اداة الاستثناء) أمّا *Âam∼A* 'as for'
  - Verbal particles such as سوف *sawfa* 'will' and قد *qad* 'may/might'
  - Negation particles such as لم *lam* 'did not' and لن *lan* 'will not'
  - The definite article ال *Al* 'the' when it appears already segmented
  - Interrogative particles such as هل *hal* 'does/is?'
  - The vocative particle يا *yA*
- **PNX** is used for all punctuation marks.

4. Syntactic Annotation Guidelines

Syntactic annotation in the dependency framework involves two types of inter-related decisions: attachment and labeling (Žabokrtský and Smrž, 2003; Habash and Rambow, 2004; Habash et al., 2007a; Smrž et al., 2008; Tounsi et al., 2009). The attachment of one word to another indicates that there is a syntactic relationship between the head (governing) word and the dependent (governed) word (and the subtree it heads). The labels, henceforth relations, specify the type of the attachment. For example, the relation, *subject*, may label the attachment of a dependent noun to a heading verb, where the noun is the subject of the verb.
In the rest of this section, we provide a top-level review of the different relations in CATiB before discussing them in the context of different syntactic constructions.

cated in parentheses): آ Ā (|), أ Â (>), ؤ ŵ (&), إ Ă (<), ئ ŷ (}), ة ħ (p), ث θ (v), ذ ð (*), ش š ($), ظ Ď (Z), ع ς (E), غ γ (g), ى ý (Y), ً ã (F), ٌ ũ (N), ٍ ĩ (K), ّ á (').

[2]PATB tokenization does not segment the definite article +ال *Al+* 'the' and neither does CATiB's. The latest version of the PATB tokenizes the future particle +س *s+* 'will'. We plan to follow their lead in the next version of CATiB.

Syntactic Relations There are eight syntactic relations that are used to label the attachments in a CATiB tree.

- **SBJ** stands for *subject*. SBJ marks the explicit syntactic subjects of verbs (active or passive) regardless of whether they appear before or after the verb (فاعل فعل او مبتدأ فاعل لخبره). SBJ also marks the subjects of nominal sentences including those headed by incomplete verbs and verb-like particles (مبتدأ لجملة اسمية بسيطة او اسم لكان او إن).
- **OBJ** stands for *object* of verbs and deverbal nouns (مفعول به) and *object* of prepositions (اسم مجرور). It is also used to mark the children of coordinating conjunctions (اسم معطوف) and subordinating conjunctions.
- **PRD** stands for *predicate*. PRD is only used to mark the complement of incomplete verbs and verb-like particles (خبر كان و خبر إن).
- **TPC** stands for *topic*. TPC has a very restricted usage. It is the subject/topic (مبتدأ) of a complex nominal sentence whose complement is a verb with a different subject. Typically there is an object pronoun that refers back to the topic.
- **IDF** stands for *idafa*. It marks the possessor in an idafa construction (مضاف اليه).
- **TMZ** stands for *tamyiz*. This relation marks the specifier in the tamyiz construction (discussed below).
- **MOD** stands for *modifier*. This is the most common relation used to mark all modifications such as adjectival modifications of nouns, adverbial modification and prepositional phrase modification of nouns and verbs.
- **—** stands for *flat*. This is a special relation used to mark multi-word structures that cannot be explained using any of the above relations. The most common case is the different parts of a proper name, e.g., a last name is in a flat relation to a first name.

Sentence Structure Arabic has three sentence structures: the verbal sentence, the nominal sentence and the complex sentence.

In the basic ***verbal sentence***, the verb is followed by a subject, object and other modifiers. The subject can be pro-dropped (conjugated) and as such may not be expressed as a separate token. The verb agrees in gender with the explicit subject but keeps a singular number. Pronominal objects follow the verb directly appearing between verb and subject. For passive verbs (VRB-PASS), SBJ is the surface subject (i.e., the underlying object). See Figures 1(a) - 1(d).

In the basic ***nominal sentence*** (also known as the equational/copular/verbless sentence), the verbless complement/predicate (الخبر) heads the topic/subject (المبتدأ). When an incomplete verb (كان واخواتها) precedes the nominal sentence, the topic/subject and complement/predicate are considered children of the incomplete verb with the relations SBJ and PRD, respectively. The same happens when a verb-like particle (إن واخواتها) precedes the nominal sentence. See Figures 1(e) - 1(g). The predicate of a nominal sentence can also be a preposition. See Figures 1(h) - 1(i).

The ***complex sentence*** is a nominal sentence whose complement/predicate (الخبر) can be a basic verbal or basic nominal sentence. The topic/subject (المبتدأ) of the complex sentence is marked as SBJ if it is the same as the subject inside the complement; otherwise it is marked as TPC. A TPC is usually referred back to using a possessive or object pronoun inside the complement; see Figures 1(j) - 1(l). When the complement is a verbal sentence, the surface word order looks like a subject-verb-object order, unlike the order of the basic verbal sentence, verb-subject-object. The verb and subject in this complex sentence agree in gender and number (as opposed to agreeing in gender only in the basic verbal sentence); compare Figure 1(a) and Figure 1(j). The complex sentence behaves like a nominal sentence when preceded by an incomplete verb or a verb-like particle. See Figures 1(m) - 1(n).

Verbal Modifiers A variety of particles can modify verbs' tense, polarity and aspect. These particles always attach under the verb with the relation MOD. See Figures 1(o) - 1(q).

Prepositional Phrases Prepositions always head their objects (OBJ) and are headed by whatever they modify (MOD). Since there are no morphological agreement restrictions on where a preposition can attach, the annotation must rely on semantics in making the attachment decision. Compare Figure 1(r) and Figure 1(s).

Sentential Modifiers In addition to prepositional phrases, adverbial nominals can also modify sentences. They are attached to the head of the sentence with the relation MOD. See Figure 1(t).

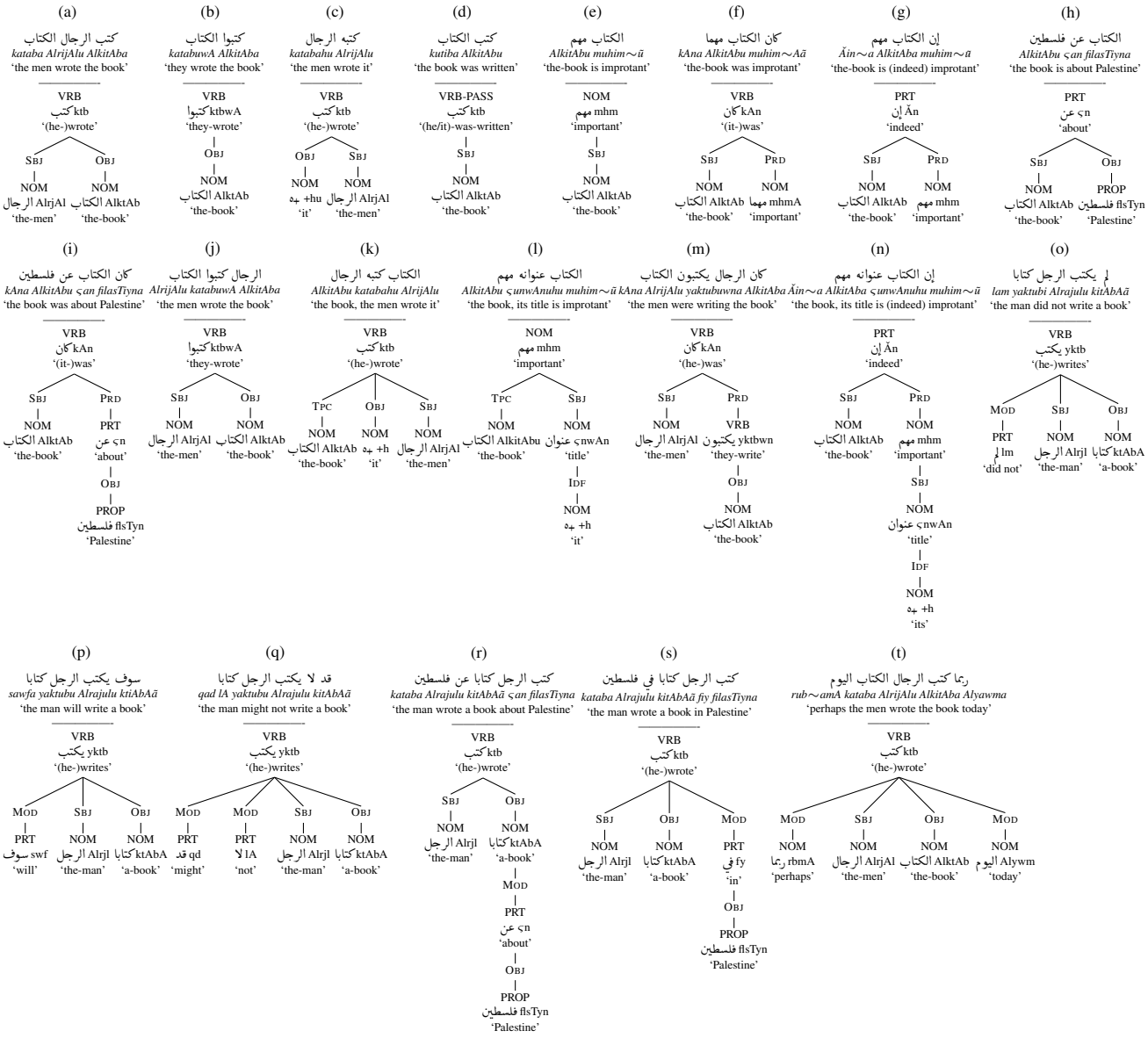
Nominal Modifiers Nominals can have three types of dependent modifiers: IDF, TMZ and MOD.

IDF is used to mark the genitive possessor (مضاف اليه) in the possessive construction, إضافة *idafa*. In addition to possession, idafa is used in various quantification constructions (with numbers 3 to 10, and 100, 1000 etc. and with general quantifiers). Idafa is also used to mark the objects of preposition-like nominal adverbs and in *clarified* adjectival modification or what is often called false idafa إضافة غير حقيقية. See Figures 2(a) - 2(f). The idafa construction can apply recursively creating what is called an *idafa chain*. See Figure 2(l). However, only one IDF is allowed per word. See Figures 2(j) - 2(k).

TMZ is used to mark the specifier in the specification construction, تمييز *tamyiz*. This construction is often used with numerals between 11 and 99 or to specify measurements. The specifier is always singular accusative. See Figure 2(g).

The most basic use of ***MOD*** is to mark adjectival modification, and the demonstrative article/pronoun, which may precede or follow the NOM it modifies; see Figures 2(h) - 2(i). Other uses of MOD include marking relative clauses and appositions (discussed below).

Figure 1: Examples of CATiB sentence-level syntactic structures



These different modification relations can combine together in different configurations. See Figures 2(m) - 2(o).

Nominal Arguments Nominals can also take OBJ arguments. This is often the case with deverbal nominals when they cannot take an argument as IDF because an IDF already exists or because the word is not in construct state morphologically. See Figures 2(j) - 2(k).

Relative Clauses Relative clauses modifying nominals can be headed by a relative pronoun (if the modified nominal is definite) or not (if the modified nominal is indefinite). In either case, the head of the relative clause is attached to the modified nominal with the relation MOD. See Figures 2(p) - 2(q).

Proper Nouns Proper nouns (PROP) are treated just as NOMs unless the structures they appear in are not explainable in terms of the basic modification relations described above. In such case, the flat relation (—) is used. See Figures 2(r) - 2(s).

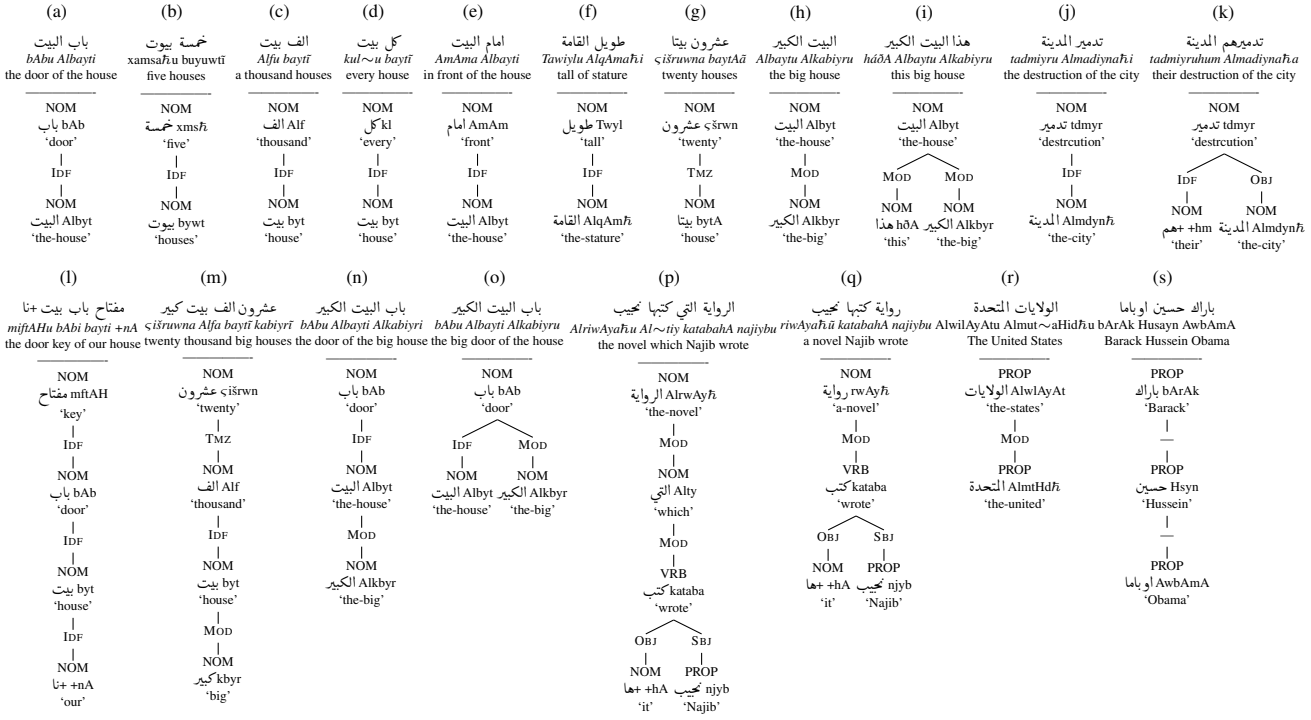
Apposition In the case of appositions (بدل), the later nominal is governed by the former nominal with the relation MOD. See Figure 3(a).

Coordination Coordination using coordinating conjunctions (حروف العطف) is annotated as follows: the head of the first joined sub-tree heads the conjunction (with relation MOD) and the head of the second joined sub-tree is headed by the conjunction (with relation OBJ). See Figures 3(b) - 3(c).

Coordinating conjunctions in Arabic, in particular the particle +و *w+* 'and', are often used as sentence-initial discourse connectives (واو ابتدائية/واو استئنافية) or interruptives (واو اعتراضية). In such cases, the conjunction is attached to the head of the sub-tree that follows it with the relation MOD. See Figures 3(d) - 3(e).

Subordination In most subordination constructions, the head of the main clause heads the subordination conjunction (with relation MOD) and the head of the subordinate

Figure 2: Examples of CATiB nominal constructions



clause is headed by the subordination conjunction (with relation OBJ). See Figures 3(f) - 3(g).

The subordinating conjunction أن *Âan* heads a subordinate VRB (with relation OBJ), but it can be attached to its head as MOD, SBJ, OBJ, TPC or IDF as appropriate. See Figures 3(h) - 3(i). Similarly, the conjunctional verb-like particle أنّ *Âan∼a* takes a SBJ and PRD as children but it can attach to its head with a variety of relations. See Figure 3(j).

Punctuation Punctuation marks are always attached as children to the words they modify with relation MOD. The general guidelines for attaching punctuation is that they attach to the highest node in the tree that explains the reason for the punctuation. For example, sentence final periods are attached to the head of the sentence; while quotation marks around quoted direct speech attach to the head of the quoted text. See Figure 3(k).

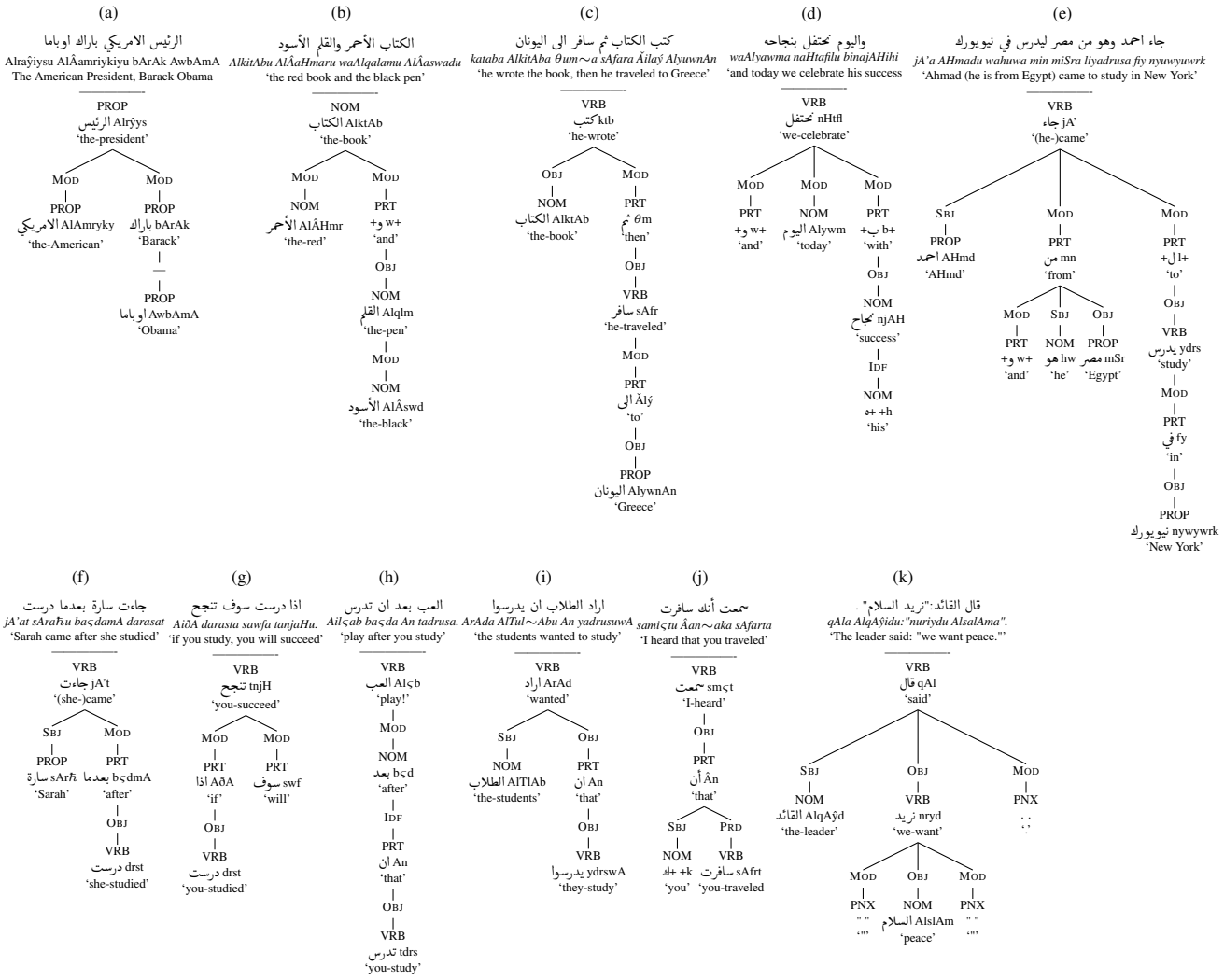
Ambiguous Attachments As in any treebanking effort, there are cases of complete ambiguity that are unresolvable through sentence/document context, e.g., the word الكبير *Alkbyr* 'the-big' in Figures 2(n) and 2(o). In such cases, we instructed the annotators to default to a low attachment as opposed to a high attachment — preferring the analysis in Figure 2(n) over Figure 2(o) in this example.

5. Comparison with PATB and PADT

When comparing PATB, PADT and CATiB, we can distinguish two high-level aspects: syntactic representation and linguistic content. In terms of syntactic representation, PABT uses phrase structure (PS) and both CATiB and PADT use dependency structure (DS). See Figure 4. PS is a tree representation in which words in a sentence appear as leaves and internal nodes are syntactic categories such as *noun phrase* (NP) or *verb phrase* (VP). DS is also a tree except that the words in the sentence are the nodes on the tree (Xia et al., 2009). In terms of linguistic content, we can further distinguish the following categories of content. In this discussion, PADT refers solely to PADT's analytical level and not PADT's deeper tectogrammatical level (unless explicitly mentioned).

Syntactic Structure PADT and CATiB annotate heads explicitly and spans of phrases/clauses implicitly; whereas PATB annotates spans explicitly and heads implicitly. PATB uses intermediate projections, such as VP, to represent certain syntactic facts. The DS treebanks, PADT and CATiB, use other devices, such as attachment labels, to represent the same facts. PADT and CATiB approach some structures differently. Here are four examples of such differences. First, in PADT the coordination conjunction heads over the different elements it coordinates as opposed to the way it is done in CATiB. See how لبنان و+ سوريا *lbnAn w+ swryA* 'Lebanon and Syria' is represented in PADT and CATiB in Figure 4. Second, in multi-word prepositions such as ب+الرغم من *b+Alrγm mn* 'in spite of', the last preposition heads the whole expression and its object in PADT; whereas in CATiB it is annotated as three words in a chain headed by the first preposition: *b+ 'in'* ← *OBJ Alrγm 'spite'* ← *MOD mn 'of'*. Third, relative pronouns in PADT are annotated as leaves in the clause they introduce; the head of the relative clause is attached directly to the modified noun; whereas in CATiB the relative pronoun heads the relative clause and is headed by the modified noun. Finally, in PADT, the subject of a verb-like particle is attached under the predicate, as opposed to being a sibling to it (as in CATiB).

Figure 3: Miscellaneous CATiB constructions



Syntactic and Semantic Functions PATB uses about 20 *dashtags* that are used for marking syntactic and semantic functions. Syntactic dashtags include -TPC and -OBJ and semantic tags includes -TMP (time) and -LOC (location). Some dashtags serve a dual semantic/syntactic purpose such as -SBJ which can mark syntactic subject of a verb and the semantic subject of a deverbal noun. PATB does not explicitly annotate dashtags in some cases such as objects of prepositions or the idafa/tamyiz constructions. These are implicitly marked through the syntactic structure. Idafa and tamyiz are identical in PATB except for the morphological case information which can be used to distinguish them. CATiB's relation labels mark syntactic function only. The use of the syntactic labels SBJ and TPC is different between CATiB and PATB. In PATB, TPC is used to mark the subject or object when they appear before the verb. Further co-indexation is used to specify the role of the TPC inside the verb phrase. See how the subject is handled in Figure 4. The subject of a verbless (non-complex) nominal sentence is marked as SBJ in both PATB and CATiB. PADT uses around 20 labels, although with different functionality from PATB and CATiB. In general, PADT analytical labels are deeper than CATiB since they are intended to be a stepping stone towards the PADT tectogrammatical level. For instance, dependents of prepositions are marked with the relation they have to the node governing the preposition (the grandparent node). For example, in Figure 4, ايلول *Aylwl* 'September' is marked Adv (Adverbial) *of the main verb* زاروا *zArwA* 'visited'. Similarly, the coordinated elements لبنان و+ سوريا *lbnAn w+ swryA* 'Lebanon and Syria' are marked as both Co (coordinated) and with their relationship to the governing verb, Obj (object). PADT does not distinguish different types of nominal modifiers, i.e. adjectives, idafa and tamyiz (in numbers) are all marked as Atr (Attribute).

Empty Pronouns Empty pronouns are annotated in PATB but not PADT nor CATiB. Verbs with no explicit subjects in CATiB (and PADT) can be assumed to pro-drop (implicit annotation).

Coreference Coreference indices are annotated in PATB for traces and explicit pronouns. PADT only annotates coreference between explicit pronouns and what they corefer with. CATiB does not annotate any coreference indices.

Word Morphology CATiB uses the same basic tokenization scheme used by PATB and PADT. As for parts-of-

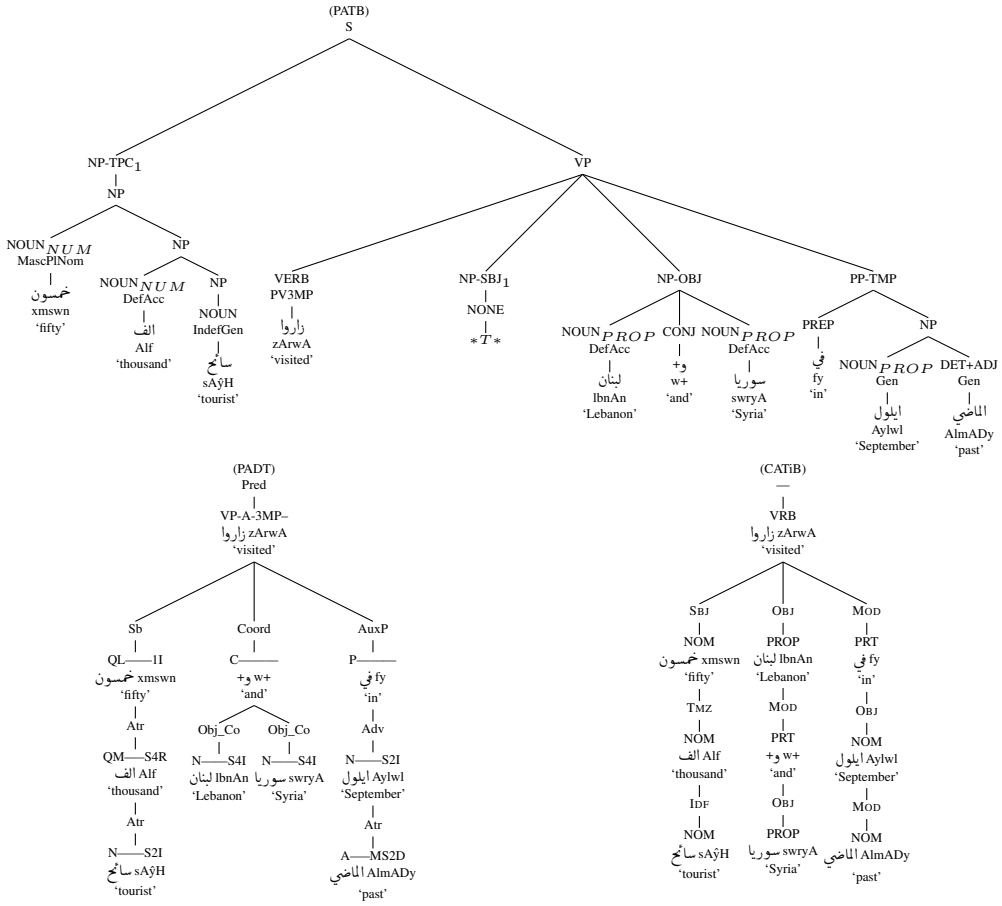


Figure 4: Comparing the phrase structure representation in the Penn Arabic Treebank (PATB) and the analytical dependency representation in the Prague Arabic Dependency Treebank (PADT) to CATiB for the sentence خمسون الف سائح زاروا لبنان وسوريا في ايلول الماضي *xmswn Alf sAŷH zArwA lbnAn wswryA fy Aylwl AlmADy* '50 thousand tourists visited Lebanon and Syria last September.'

speech, PATB uses over 400 tags specifying every aspect of Arabic word morphology such as definiteness, gender, number, person, mood, voice and case. PADT morphology is more complex than PATB. For instance, it makes more sophisticated distinctions on nominal and adjectival definiteness/state, number, and gender. In contrast, CATiB uses six POS tags only. It is important to point out that in most Arabic parsing work, a much smaller POS tagset is used, reducing the 400 or so tags in PATB to a set between 20 and 40 tags. We were able to reproduce one of these tagsets (Kulick et al., 2006) automatically at 98.5% accuracy using features from the annotated trees. A majority of the remaining errors are confusion in distinguishing nominals (noun/adjective/adverb). Details of this result will be presented in a future publication. Some of the rich morphology information not included in reduced POS tagsets, such as nominal case, can also be retrieved from the tree structure because they are defined syntactically (Habash et al., 2007a).

Despite the many differences, conversion between these different representation can be done with a good degree of success given that the information is available in the tree although represented differently. Since CATiB has less content than PATB and PADT, it is perhaps much easier to convert from these two representations into CATiB's than the other way around.

6. CATiB Package

Data Sets CATiB annotated data is taken from the following LDC-provided resources:[3] LDC2007E46, LDC2007E87, GALE-DEV07, MT05 test set, MT06 test set, and a small portion of PATB (part 3). These datasets are 2004-2007 newswire feeds collected from different news agencies and news papers, such as Agence France Presse, Xinhua, Al-Hayat, Al-Asharq Al-Awsat, Al-Quds Al-Arabi, An-Nahar, Al-Ahram and As-Sabah. The CATiB-annotated PATB portion was extracted from An-Nahar news articles from 2002. Headlines, datelines and bylines are not annotated and some sentences are excluded for excessive (>300 tokens) length and formatting problems. Over 273K tokens (228K words, 7,121 trees) of data were annotated, not counting duplications for computing inter-annotator agreement. In addition, the PATB part 1, part 2 and part 3 data is automatically converted into CATiB representation. This converted data contributes an additional 735K tokens (613K words, 24,198 trees). Collectively, the CATiB version 1.0 release contains over 1M tokens (841K words, 31,319 trees),

[3] http://www.ldc.upenn.edu/

including annotated and converted data. CATiB is now available through the LDC (LDC2009E06).

CATiB Release Components The release consists of a large collection of XML files. For each document, the XML includes tags to hold the original LDC document id number, the raw source sentence, the translation (if available), a tokenized version of the sentence, and a dependency tree. The dependency is represented in two formats: (a.) as a list of 5-tuple per word specifying word position, word form, POS tag, parent position, and relation and (b.) as a phrase-structure-like tree with explicit heads, spans, POS tags and relations. All Arabic script is UTF-8 encoded.

7. Future Outlook

We would like to extend CATiB annotation to handle genres other than newswire, e.g., broadcast news/conversation, speech transcripts, web text, poetry, etc. We also would like to consider non-Modern Standard Arabic texts, including both Quranic/classical and dialectal Arabic texts.

8. Acknowledgements

We would like to thank all the annotators who worked very hard on creating CATiB: Mayss Bajbouj Kinjawi, Jamila Kamal El-Gizuli, Iman Issa, Guzide Kobati, Sahar Masri Jendi and Mohamed Nasr. We would like to thank Owen Rambow and Otakar Smrž for helpful discussions. We would like to thank the members of the PATB team, Mohamed Maamouri, Ann Bies and Seth Kulick for the excellent resources they have created, the PATB and its manual, which we used extensively. We would like to thank the GALE Banks committee members for their support and guidance. This work has been supported by Defense Advanced Research Projects Agency Contract No. HR0011-08-C-0110. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

9. References

Tim Buckwalter. 2002. Buckwalter Arabic morphological analyzer version 1.0.

Nizar Habash and Owen Rambow. 2004. Extracting a Tree Adjoining Grammar from the Penn Arabic Treebank. In *Proceedings of Traitement Automatique du Langage Naturel (TALN-04)*, pages 277–284. Fez, Morocco.

Nizar Habash and Owen Rambow. 2007. Arabic Diacritization through Full Morphological Tagging. In *Proceedings of the 8th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL07)*.

Nizar Habash and Ryan Roth. 2009. CATiB: The Columbia Arabic Treebank. Technical Report CCLS-09-01, Center for Computational Learning Systems, Columbia University.

Nizar Habash, Ryan Gabbard, Owen Rambow, Seth Kulick, and Mitch Marcus. 2007a. Determining case in Arabic: Learning complex linguistic behavior requires complex linguistic features. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1084–1092.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007b. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash, Reem Faraj, and Owen Rambow. 2009. Columbia Arabic Treebank Annotation Manual. http://www1.ccls.columbia.edu/ CATiB/.

Seth Kulick, Ryan Gabbard, and Mitch Marcus. 2006. Parsing the arabic treebank: Analysis and improvements. In *Proceedings of the Treebanks and Linguistic Theories Conference*, pages 31–42, Prague, Czech Republic.

Mohamed Maamouri, Ann Bies, and Tim Buckwalter. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Mohamed Maamouri, Ann Bies, and Seth Kulick. 2009. Creating a Methodology for Large-Scale Correction of Treebank Annotation: The Case of the Arabic Treebank. In *Proceedings of MEDAR International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of ACL-08: HLT, Short Papers*, pages 117–120, Columbus, Ohio, June. Association for Computational Linguistics.

Otakar Smrž and Jan Hajič. 2006. The Other Arabic Treebank: Prague Dependencies and Functions. In Ali Farghaly, editor, *Arabic Computational Linguistics: Current Implementations*. CSLI Publications.

Otakar Smrž, Viktor Bielický, Iveta Kouřilová, Jakub Kráčmar, Jan Hajič, and Petr Zemánek. 2008. Prague arabic dependency treebank: A word on the million words. In *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008)*, pages 16–23, Marrakech, Morocco.

Lamia Tounsi, Mohammed Attia, and Josef van Genabith. 2009. Automatic treebank-based acquisition of Arabic LFG dependency structures. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 45–52, Athens, Greece.

Fei Xia, Owen Rambow, Rajesh Bhatt, Martha Palmer, and Dipti Misra Sharma. 2009. Towards a Multi-Representational Treebank. In *Proceedings of Treebanks and Linguistic Theories (TLT 7)*, Groningen, Netherlands.

Zdeněk Žabokrtský and Otakar Smrž. 2003. Arabic syntactic trees: from constituency to dependency. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL'03) – Research Notes*, Budapest, Hungary.