Explicit Fine Grained Syntactic and Semantic Annotation of the Idafa Construction in Arabic

Abdelati Hawwari, [+]Mohammed Attia, Mahmoud Ghoneim, Mona Diab

Department of Computer Science, George Washington University
[+]Google Inc.
{abhawwari, mghoneim, mtdiab}@gwu.edu; attia@google.com

Abstract

Idafa in traditional Arabic grammar is an umbrella construction that covers several phenomena including what is expressed in English as noun-noun compounds and Saxon & Norman genitives. Additionally, *Idafa* participates in some other constructions, such as quantifiers, quasi-prepositions, and adjectives. Identifying the various types of the Idafa construction (IC) is of importance to Natural Language Processing (NLP) applications. Noun-Noun compounds exhibit special behaviour in most languages impacting their semantic interpretation. Hence distinguishing them could have an impact on downstream NLP applications. The most comprehensive computational syntactic representation of the Arabic language is found in the LDC Arabic Treebank (ATB). Despite its coverage, ICs are not explicitly labeled in the ATB and furthermore, there is no clear distinction between ICs of noun-noun relations and other traditional ICs. Hence, we devise a detailed syntactic and semantic typification process of the IC phenomenon in Arabic. We target the ATB as a platform for this classification. We render the ATB annotated with explicit IC labels in addition to further semantic characterization which is useful for syntactic, semantic and cross language processing. Our typification of IC comprises 3 main syntactic IC types: False Idafas (FIC), Grammatical Idafas (GIC), and True Idafas (TIC), which are further divided into 10 syntactic subclasses. The TIC group is further classified into semantic relations. We devise a method for automatic IC labeling and compare its yield against the CATiB Treebank. Our evaluation shows that we achieve the same level of accuracy, but with the additional fine-grained classification into the various syntactic and semantic types.

Keywords: Arabic, construct state, Idafa, annotation, Treebank, syntax, semantics

1. Introduction

1.1. Idafa in Arabic

Idafa is an Arabic term that means "annexation" or "addition". In Arabic grammar, Idafa is a construction that is made up of two nominal parts (nouns, adjectives, proper nouns), where the whole construction serves as a single syntactic unit. The first part, in Arabic "*mudaf*" (MDF), is an indefinite noun and the second one, "*mudaaf elayh*" (MDFE), could either be definite or indefinite. Typically, in Idafa constructions (IC) the MDFE defines or specifies the MDF (Boujelben et al., 2011). Habash et al. (2007) highlight the role IC plays in syntactic case realization for nominals in Arabic. Syntactic case marking depends on whether the nominal is indefinite, namely marked with the nunation/tanween diacritic typically expressed as a word final *an, un, in*, or if the nominal is definite through agglutinating the definite article *Al+* as a prefix, or is definite through IC. The case on MDF is sometimes referred to as the construct state. In principle, IC could be recursive with no specific bounds on the number of embeddings or nestings.

IC in Arabic is practically a wide construction that covers many linguistic phenomena related to noun compounding. In traditional grammar, IC covers relations between quantifiers and noun phrases they scope over, numbers and their head nouns, proper nouns, verbal nouns and their subjects/objects, etc.

There is no exact equivalent to IC in English. When studying Arabic *Idafa*, it is a common misconception to compare it to English genitive construction. In fact, Arabic Idafa is used for what may be expressed in English as noun-noun compounds, e.g. *bus stop*; the Saxon genitive, e.g. *student's book*; the Norman genitive, e.g. *end of the day*; in addition to many other constructions, such as quantifiers, prepositions, and adjectival phrases. Accordingly, it is conventionally preferred to use the Arabic term *Idafa*.

It is worth noting that the linguistic phenomenon of IC in Arabic has a counterpart in some related languages, such as Persian, aka *Ezafe*, and Hebrew, aka *smikhut* (SC). Hence, identifying and clearly classifying the IC in Arabic may have significant implications on the characterization of ICs in these languages.

The IC is an important phrase structure in Arabic due to its ubiquity and the variability of its syntactic compositions and semantic relations. In a study of the LDC Arabic Treebank (ATB) (Maamouri et al. 2004), Green and Manning (2010) state that the average number of ICs per sentence is 3.06 and it occurs in 84.3% of the trees in their development set. Gabbard and Kulick (2008) show that there is significant attachment ambiguity associated with IC, and reveal that one-level ICs make up more than 75% of NP ICs in the ATB, while recursive ICs make up the rest, 25%, with a maximum of 6 levels of embedding. Gabbard and Kulick (2008) also point out that 40.9% of ICs are modified by one or more adjectives. By comparison, Moldovan et al. (2004) have shown that 60% of complex nominals in the Wall Street Journal Corpus are composed of Saxon genitives, 8% are noun-noun compounds, and 32% are Norman genitives.

We postulate that a clear explicit labeling which makes distinction between different types of ICs could be of crucial significance for several downstream applications that care about semantic compositionality such as machine translation, parsing, multiword expression detection, and information extraction such as named entity recognition. Moreover, having an exhaustive labeling of ICs and their various types would allow for more precise empirical studies of Arabic linguistic phenomena on a large scale.

The ATB serves as one of the main reference points on large-scale Arabic syntactic representations for computational processing. It is also used as the gold standard in Arabic NLP, making it a resource of significant importance. Implicitly, ICs are present in the ATB as an NP headed by a NOUN and an NP complement (Maamouri et al., 2008). However, the ATB does not attempt to distinguish ICs explicitly nor differentiate between the different types of ICs. Accordingly, no distinction is made between noun-noun and quantifier-noun constructions. The phrases كِتابُ النَّحْوِ *kitAbu Al-naHowi* [1] "grammar book" and كُلُّ المَواضيعِ *kul~u Al-mawADiyEi* "all topics" are rendered with the same syntactic realization, i.e. both are simply labeled (NP (NN NP)) where the words for *book* and *all* are labeled the same way as NN. The only attempt we know of to explicitly label ICs on a large scale was in the context of the CATiB dependency annotation effort in which they use the IDF tag, as one of 8 grammatical relation labels (Habash and Roth, 2009).

Accordingly, in this paper, we present a typification of the various coarse syntactic constructions of (NP (NN NP)) and (NP (JJ NP)) present in the ATB, corresponding to the syntactic IC linguistic phenomenon, into their various IC types based on the true POS categories of their heads and complements. We present an automatic extraction and classification process which yields three main types that are further divided into 10 syntactic IC subtypes. Furthermore, we present our semantic framework for the classification of the nominal constructions of the ICs. We limit our investigation to Modern Standard Arabic in this paper.

1 All the transliteration in this paper is presented using the Buckwalter encoding system www.qamus.com

1.2. Idafa in other languages

The linguistic phenomenon of *Idafa* in Arabic has counterparts in four languages: Persian, Urdu, Kurdish (known as *Esafe* in the three languages) and Hebrew (known as *smikhut*). Here we will briefly introduce Idafa in these languages and show how they compare and contrast to Arabic.

Persian Ezafe (EZ) as described by Megerdoomian (2000) shares a lot of the characteristics and properties of Idafa in Arabic. For example, like Arabic, Persian is head-initial, and genitive nouns and pronouns follow the head noun. Pronouns and proper names mark the boundary of the Noun Phrase (i.e., no further NP attachment is allowed to the right of pronouns and proper names). Complex or recursive Ezafe is a series of consecutive nouns, with each noun forming a simple binary Ezafe. Persian Ezafe is expressed through a suffix "e" which is usually pronounced as a short vowel and typically underspecified in written text. For example, */ketAb-e dAryush/*, literally *book-EZ Dariush*, meaning "Dariush's book"

Similarly, Idafa in Arabic triggers a similar morpho-syntactic process referred to as construct state. Habash et al., (2007) points out that Idafa plays an important role in case realization for nouns in Arabic. Case marking depends on whether the nominal is indefinite, i.e., receiving nunation (tanween), definite through the definite article Al+ or definite through Idafa. When indefinite nouns appear outside of Idafa they are marked with a tanween diacritic. In IC this diacritic is reduced. Moreover, for example, masculine sound plural nouns lose their word final nuun 'n'.

Lotfi (2014) analyzed Persian Ezafe as lending support to his Integrity Hypothesis which assumes an interface between syntactic word order and semantic relations. He states that when an entity is tightly integrated into another, the intrinsically salient entity will appear as the head (as in whole-part, possessor-possessed, actor-location and functor variable relationships), but when the integration is loose either entity may be extrinsically more salient, and can surface as the head (as in container-content, producer-product, group-member, and head-body relationships).

This construction is also found in Urdu, an Indo-Aryan language, as in the following example /talib-e Elm/, literally seeker.masc.sg-EZ knowledge.masc.sg, meaning "student". (Bögel and Butt, 2013) analyzed Urdu Ezafe as reflecting an interplay between phonology morphology and syntax.

Ezafe, or *Ezafet*, is also found in Kurdish, an Indo-Iranian language, as in the following example (Strunk, 2003), /mal-a vi mer-e/, literally *house.fem.sg-EZ this.dem.sg.obl man.masc.sg.obl*, meaning "this man's house".

Samvelian (2007) points out that adjectives and some prepositions may take the Ezafe in Kurdish and Persian (as in Arabic) since adjectives behave like nouns in many respects. Moreover, prepositions occurring with the Ezafe are not in fact 'true' prepositions, but display a range of nominal properties.

A similar construction is found in Hebrew and is called *smikhut* (Shilon et al., 2012) in which, like Arabic, the head of the structure is the first noun, which typically determines the number and gender agreement features, for example, /sfr h+ild/, literally *book.indef the+boy.def*, meaning "The boy's book".

Unlike Arabic (more like Arabic dialects in fact), Hebrew alternatively uses a possessive preposition šl 'of' as in the example /h+sfr šl h+ild/ literally *the+book.def of the+boy.def*, meaning "the boy's book". Hebrew also uses double genitives in which the antecedent noun is followed both by a cliticized possessive pronoun and the genitive marker šl as in the following example: /sfr+w šl h+ild/, literally *book+his of the+boy.def*, meaning "The boy's book".

2. Idafa Construction Typification

A general misconception about ICs in Arabic is that they are composed of two nouns annexed to each other. In fact, the first part of Idafa, the MDF, can be an adjective, a function noun (preposition, quantifier, etc. functioning as a noun) or a true noun. Accordingly, there are three types of ICs in the ATB that have the syntactic realization of (NP (NN NP)) and (NP (JJ NP)). These three types are False Idafa Constructions (FIC), Grammatical Idafa Constructions (GIC), and True Idafa Constructions (TIC).

2.1. False Idafa Constructions (FIC)

One way of classifying IC in traditional Arabic grammar is to distinguish between *real Idafa* and *false Idafa*. False Idafa is used to refer to Idafa headed by adjectives. The frequency of FIC in the ATB is 1.4%. There are three types of FICs, depending on whether they are headed by a qualitative adjective, an elative adjective, or the construction is nominalized.

2.1.1. Qualitative FIC

Qualitative FIC is where an adjective goes in the MDF position and the whole phrase modifies or predicates a preceding noun, examples: بَعِيدُ المَنال, *baEiydu Al manAl*, "hard to reach"; بالِغُ الخُطُورَةِ, *bAligu Al-xutuwrapi*, "extremely dangerous". The adjective agrees with the preceding noun in number, gender and definiteness, e.g. الأَهْدافُ البَعيدَةُ المَنال, *Al->ohodAfu Al-baEiydapu Al-manAli,* "hard-to-reach goals".

2.1.2. Elative FIC

Elative false Idafa is an adjective in the elative (تفضيل tafodiyl) form that goes in the MDF position, and is invariably in the singular masculine form, examples: أَقْرَبُ فُرْصَةٍ, *>aqorabu furosapK*, "nearest opportunity"; أعْلَى دَرَجَةٍ, *>aEolaY darajapK*, "higher degree". It is to be noted that the morpho-syntactic difference between qualitative and elative FIC is that no agreement in number or gender is displayed on the adjectives in the case of the elative FIC.

2.1.3. Nominalized FIC

Nominalized FIC is an adjective (usually in the masculine, plural form) that goes in the MDF position followed by a noun and the adjective itself behaves like a noun, i.e. it does not modify or predicate, but rather tends to occupy the syntactic positions of subject, object or object of preposition for example: كِبارُ التُّجّارِ, *kibAru Al-tuj~Ari*, "leading traders"; مَحْدودُو الدخْل, *maHoduwduw Al-daxoli*, "limited-income people".

2.2. Grammatical Idafa Constructions (GIC)

These comprise several syntactic phenomena that do not include the typical noun-noun relationships. In this class of Idafa, the first element, MDF, can be a quasi-preposition, quantifier, number, verbal masdar, or proper noun. It also includes cases where the second element of Idafa, MDFE, is a verbal masdar or complement clause. As one of the two constituent parts in Idafa loses it nominal function, Idafa becomes purely grammatical and no semantic relations can be induced. Most of these cases are deterministically identified.

2.2.1. Quasi-Prepositions IC

Quasi-prepositions, or prepositionals, are nominal elements that behave like prepositions, such as: مَعَ, *maEa*, "with"; بَعْدَ, *baEoda*, "after"; فَوْقَ, *fawoqa,* "above", however they function as heads in IC since they lie on the border between nouns and prepositions. Quasi-prepositions in the ATB include the following list: أَمامَ *>amAma* "in front of", أَثْناءَ *>avonA'a* "during", ضِدَّ *Did~a* "against", ضِمْنَ *Dimona* "among", إِثْرَ *<ivora* "after", إزاءَ *<izA'a* "towards", عَصْرَ *EaSora* "the evening of", عَبْرَ *Eabora* "across", عِوَضَ *EiwaDa* "instead of", حَوْلَ *Hawola* "around", حِيالَ *HiyAla* "towards", صَباحَ *SabAHa* "the morning of", بَدَلَ *badala* "instead of", بَيْنَ *bayona* "between", دُونَ *duwna* "without", فَوْقَ *fawoqa* "above", غَيْرَ *gayora* "not including", مَساءَ *masA'a* "the night of", مِثْلَ *mivola* "such as", قَبْلَ *qabola* "before", قرب *quroba* "near", رَغْمَ *ragoma* "despite", تَحْتَ *taHota* "under", تِجَاهَ *tujAha* "towards", وَرَاءَ *warA'a* "behind", وَسْطَ *wasoTa* "midst", وِفْقَ *wifoqa* "according to", خارِجَ *xArija* "outside of", and خِلالَ *xilAla* "through". In Arabic traditional grammar they are called الأسْماء المُلازِمة للإضافَة *Al->asomA' Al-mulAzimap l-l-<iDafap* "nouns adhering to Idafa" and المَفعول فِيه *Al-mafEuwl fiyhi* "locative/temporal accusatives".

Quasi-prepositions are distinguished by the following characteristics: a) derivationally, they are nouns, and in some contexts they retain some nominal properties, such as taking the definite article ال e.g. الأمام, *Al->amAmi,* "the front", and they can be preceded by primitive prepositions, e.g. بِرَغْم *biragomi* "despite"; b) Morpho-syntactically, they are invariably in the accusative case; c) they do not inflect for number or gender; d) they have a different syntactic distribution than regular nouns as they never occupy the syntactic positions of subject, or object; and, e) they are mostly either locative or temporal. This class of words is annotated as NOUN and not PREP in the ATB. Since the function of Idafa is either defining or specifying, neither of these functions are served with quasi-prepositions, hence the need for a clear explicit distinction between these constructions and true idafas.

2.2.2. Quantifier IC

This is the case where the MDF is a quantifier. Partitive modification relays that X refers to a part or the whole of Y .e.g. بَعْضُ الناس *baEoDu Al-nAsi* "some people", كُلُّ الأعْضاء *kul~u Al->aEoDA'i* "all members", مُعْظَمُ الزُّوَّار *muEoZamu Al-zuw~Ari* "most visitors". Quantifiers are annotated in the ATB as NOUN_QUANT, and they include the following set: أغلب *>agolab* "most", أكْثَر *>akovar* "most", ضِعْف *DiEof* "double", بِضْع *biDoE* "few", غالِبِيَّة *gAlibiy~ap* "majority of", جَمِيع *jamiyE* "all", نِصْف *niSof* "half", ثُلُث *vuluv* "third". Quantifiers are characterized by the following: a) they don't take gender or number morphology; b) they behave like prenominal adjectival modifiers; c) they are inflected for case and superficially function as the head of the phrase; and, d) they do not control agreement, but agreement is rather determined by the number and gender of the noun that follows (the real heads), for example: كُلُّ الطلَّابِ حَضَرُوا, *kul~u Al-Tul~Abi.masc.pl HaDaruwA.masc.pl*, "All (male) students attended"; كُلُّ الطالباتِ حَضَرْنَ, *kul~u Al-TAlibAti.fem.pl HaDarona.fem.pl*, "All (female) students attended". This example illustrates also that *kul~u* takes no gender or number morphemes as mentioned in (a).

2.2.3. Number IC

This is the case where the MDF is a number. Arabic has a complicated number system. The number superficially functions as the head, while the modified noun is expressed as MDFE (for the numbers from three to ten), or a specifier (tamyeez) (for all the numbers above ten). For the numbers from three to ten, the relation between the number and the modified noun is controlled by an anti-agreement, where a number takes on the opposite gender morpheme of the modified noun, as illustrated in this example from the ATB: (NN ثَلاثَة *valAvapu*) (NP (NN رِجالِ rijAlK "men").

2.2.4. Verbal Masdar IC

Verbal masdars in Arabic are verbal nouns, which are similar to English gerunds. Similar to English, they exhibit the same type of ambiguity, whether the nominal or verbal function is intended. In English, for example, the gerund *swimming* can be nominal in *Swimming is a good sport* or verbal as in *Swimming in the river is risky*. In Arabic, a masdar is a type of nominal inflection that can retain the verbal function. For example, the verb كتب *kataba* "to write" can be inflected into the masdar as كتابة *kitAbap* "writing" and it can have a nominal function as in يُحِبُّ الكِتابةَ التحْلِيلِيَّة *yuHib~u Al-kitAbapa Al-taHoliyliy~api* "He likes analytical writing", or verbal as in بَعْدَ كِتابَتِهِ المَقالَ *baEoda kitAbatihi Al-maqAla* "after his writing (of) the article". In the latter example, the verbal function is the intended reading since "Al-maqAla" takes on an accusative case indicating a syntactic object role. Disambiguating the masdar becomes harder when there is nothing to indicate either reading, namely the nominal reading (such as having an adjectival modifier) or the verbal reading (such as having an object in the accusative case), as in يُحِبُّ قِراءَةَ الكُتُبِ *yuhib~u qirA'apa Al-kutubi* "he likes to read/reading books." In the ATB annotation, no particular effort was made to resolve ambiguous cases with masdar, and they were only marked as verbal when they govern an object in the accusative case, for example: (VP (NN.VN لِقاء *liqA}+i*) (NP-SBJ (PRON ه *hi*)) (NP-OBJ (NN_ACC الرئيسَ *Al+ra}iys+a*). As the ATB analysis shows, these verbal masdars are treated as verbs and they cannot participate in nominal relations as their nominal status is weak.

2.2.5. Proper Noun IC

The inclusion of this type among Idafa classes is an artifact of the convention chosen by the LDC in their ATB annotation scheme. We believe the more appropriate syntactic representation of proper nouns is to express them in a flat structure, however they are expressed in the ATB as a NN followed by an NP sharing the same NP mother node, as typical of IC. For example, the compound noun رِيال مَدْرِيد *riyAl madriyd* "Real Madrid" is marked as: NP (NN ريال *riyal* (NP مدريد *madriyd*)). It is to be noted that in the ATB, persons' first-last name constructions are annotated as flat structures (correctly) disqualifying them from the Idafa treatment, e.g. وَليد جُنْبلاط *waliyd junoblAT* "Walid Jumblatt" is annotated as follows: (NP (NOUN_PROP وَليد waliyd) (NOUN_PROP جُنْبلاط junoblAT)).

2.2.6. Verbal Noun Complement IC

In this case the MDFE is a verbal noun. Examples of this type are: إمْكانَ تَكَبُّدِهِمْ خَسائِرَ *<imokAna takab~udihim xasA}ira* "the possibility of them suffering losses", and نَتِيجَةَ سَدِّهِ قِطْعَة *natiyjapa sad~ihi qiToE+ap+F* "as a result of his blocking a piece".

2.2.7. SBAR Complement IC

In this case the MDFE is an SBAR construction headed by a complementizer. Examples of this type are: حَدِّ أَنَّكَ تَعْجَزُ *Had~+i >an~a-ka ta+Eojaz+u* "the limit that you are unable", and مَعْنى أنْ تُعْلِنَ اللجْنَةُ *maEonaY >ano tu+Eolin+a Al+lajon+ap+u* "the meaning that the committee declares".

2.3. True Idafas (TIC)

These comprise ICs where both the MDF and MDFE are true nouns and not functioning grammatically as something else as described above) and the modifying noun (MDFE) *specifies* (Takhsees) or defines (Tareef) the head noun. There is a fine distinction between 'specifies' and 'defines': as specifying is a kind of qualification that stops short of granting the qualified noun the status of being definite, while defining makes the construction fully definite. Definiteness is a morpho-syntactic feature in Arabic that is involved in agreement and other constraints in the sentence structure. The component parts in this group are involved in typical semantic relations of compound nouns, which will be discussed in details in Section 5.

2.3.1. Specifying or Takhsees IC

Takhsees (Idafa of specification) is a noun-noun compound where the MDFE has an indefinite case, namely missing the definite article Al and not participating by itself in another succeeding defining IC. The construction usually indicates collocations. Many nominal MWEs in Arabic use this type of Idafa in their internal structure. They convey the sense that: there are different types of X, and Y is one of them. Examples of this type are اجْتِماعُ عَمَلٍ *{ijotimaEu EamalK* "business meeting", مَشروعُ قانونٍ *ma$oruwEu qAnuwnK* "lit. proposal of a law; motion", and نَزْلَةُ بَرْدٍ *nazolapu barodK* "lit: fit of cold; flu". In the ATB, this kind of Idafa has the following annotation: (NP (NN+CASE_DEF_NOM اجتماعُ *{ijotimaEu*)(NP (NN+CASE_INDEF_GEN عَمَلٍ *EamalK*))).

2.3.2. Defining or Tareef IC

Tareef (defining) IC is the default genitive construction where the MDF is an indefinite noun and the MDFE is a definite noun, and the phrase acquires the status of a fully definite NP. Examples of this type are رئيسُ اللجنةِ *ra}iysu Al-lajonapi* "head of the committee", وزيرُ الداخليةِ *waziyru Al-dAxiliy~api* "minister of the interior" and وقْفُ النارِ *waqofu Al-nAri* "ceasefire". The ATB annotation is as follows: (NP (NN+CASE_DEF_NOM رئيسُ *ra}iysu*)(NP (NN+CASE_INDEF_GEN اللجنةِ *Al-lajonapi*))).

3. Automatic Syntactic Typification of IC

We develop a tool for the automatic extraction and typification of ICs in the ATB (Maamouri et al., 2010) based on syntactic, morphological and lexical clues. First, we deterministically assume that all syntactic constructions of the form (NP (NN/JJ-def NP-gen-def/indef) to be candidate ICs. These amount to 56,586 of a total possible 64,029 (NN/JJ NP) constructions. Then, we look closely at the morphological marker on the NN, which is the MDF portion of the IC. If it does not bear the definite article ال Al "the", then it is considered a valid potential MDF. Lexical clues are also used in the typification of Idafa, such as using the closed list of quasi-prepositions and quantifiers.

The MDFE portion can be definite or indefinite or a complementizer comprising a verbal noun construction or an SBAR. The average number of general IC per sentence is 4.4 (number of sentences: 12,628; and number of IC: 56,901). Table 1 shows the frequency distribution of IC types and their subtypes in the ATB. We classified the identified ICs based on the MDF and MDFE using pattern matches and a rule based system.

IC Type	*Subtype*	*Subtype Ratio %*	*IC Type Ratio%*
FIC	ADJ	1.44	1.44
GIC	Quasi Prepositional	12.13	23.73
	Quantifier	3.52	
	Number	3.31	
	Verbal Masdar	2.32	
	Verbal Noun Complement	0.61	
	SBAR Complement	1.66	
	Proper Noun	0.23	
TIC	Takhsees (designation)	4.20	77.31
	Tareef (definition)	73.11	

Table 1: Types and Frequency of IC in the ATB.

4. Evaluating IC Extraction Algorithm

We compare our extraction algorithm to the CATiB schema (Habash and Roth, 2009). CATiB representation and annotation procedure includes explicit annotation of the traditional Idafa as a separate dependency relation named "IDF" which is defined as the relation between the possessor [dependent] (corresponding to our MDFE definition) to the possessed [head] in the Idafa nominal construction (corresponding to our MDF definition). We mark the CATiB dependent, and all its dependents, as MDFE, and the head as MDF. We compare the accuracy of our extraction of IC to the CATiB schema. CATiB system identified 57,700 Idafa constructs, while our system identified 56,741 IC. There are 56,268 full matching constructs and 117 constructs that have at least

one part mismatch. We exclusively identifies 356 constructs that are mostly "verbal masdar" ICs. On the other hand, CATiB exclusively identifies 1,315 ICs. Some of these cases are not Idafa or represent problematic cases that don't have wide linguistic agreement. For example, there are 435 instances with MDF containing the definite article *Al+*. While this is not acceptable from a prescriptive classical Arabic grammar perspective, we found these construct in use in some genre like newswire and discussion fora (e.g. الايطالية الصنع , Al<yTAlyp AlSnE, Italian made").

	Precision	*Recall*	
Total IC	99.17%	97.52%	Full
MDF	99.37%	97.55%	MDF
MDFE	99.25%	97.43%	MDFE

Table 2: Comparing our IC extraction to CATiB.

5. TIC Semantic Classification

Interpretation of noun compounds is an important problem for many NLP applications. Many attempts have been made for the annotation of semantic relations involved in these constructions, yet there has been considerable disagreement regarding the number and nature of the relations (Tratz and Hovy, 2010). We review the classification/annotation schemes in the literature and adopt a taxonomy that most suits the nature of the task and data in question, and then show the results of applying this taxonomy to a subset of our data. Accordingly, we classify the TIC group, namely noun–noun compounds, into several semantic types.

5.1. Previous work on compound noun semantic relations

5.1.1. Theoretical Studies:

The first taxonomy of noun compounds can be traced back to Lee (1960) who assumed 9 classes: *subject–predicate, subject–middle object, subject–verb, subject–object, verb–object, verb–prepositional object, object–prepositional object,* and*, proper nouns*. The disadvantage of this classification is that it is based purely on the underlying syntactic relationships between constituents, rather than semantic conceptual relations.

Downing (1977) accounted for the semantic relationships between constituents in noun compounds with 12 relationships: *whole–part, half–half, part–whole, composition, comparison, time, place, source, product, user, purpose,* and*, occupation*. This taxonomy was the foundation for subsequent semantic typifications that rely more on semantic concepts, rather than lexical or grammatical relations.

Levi (1978) created 9 different classes from the generative semantic perspective using a mix of semantic concepts and prepositions: *cause, have, make, use, be, in, for, from, about*.

Warren (1978) classified the relations of noun compounds into 12 different types: *source–result, copular, resemblance, whole–part, part–whole, size–whole, goal–object, place–object, time–object, origin–object, purpose, activity–actor*. This taxonomy significantly overlaps with that of Downing (1977).

A more recent taxonomy produced by (Lauer, 1995) is based purely on prepositional paraphrases with 8 relationships: *of, for, in, at, on, from, with, about*. We argue against this taxonomy as will be discussed in the next section when justifying for our own classification methodology.

Linh (2010) devised a classification scheme of the semantic relationships between compound nouns in the technical domain (electrical engineering), creating 8 classes: *location, purpose, measure, representation, source, operation, object, structure*. The purpose of this study, however, was not to serve NLP applications, but rather to teach technical vocabulary to engineering students who study English as a second language.

Gadalla and Abdel-Hamid (2002) provide a contrastive study between the genitive constructions in Arabic and English. They provide semantic classification of genitive constructions in both languages based on 13 types: *possessive, subjective, objective, descriptive, genitive of origin, measure, place, time, relation, substance and naming, partitive and confirmation.*

5.1.2. Empirical Studies

Vanderwende (1994) described an algorithm for automatically interpreting noun sequences in unrestricted text classifying them into 13 different semantic relations: *subject, object, locative, time, possessive, whole-part, part-whole, equative, instrument, purpose, material, causes, and caused-by*.

Barker and Szpakowicz (1998) developed a semi-automatic system for identifying semantic relationships in noun phrases depending on lexical and syntactic clues. Their analyzer used 20 different relations: *agent, beneficiary, cause, container, content, destination, equative, instrument, located, location, material, object, possessor, product, property, purpose, result, source, time and topic*. Their experiments on English technical texts correctly identifies 60-70% of relationships automatically.

Moldovan et al (2004) assumed a set of 35 semantic relations in noun phrases (including noun compounds, Saxon genitive, Norman genitive, nouns modified by adjectives derived from nouns, and nouns qualified by prepositions), but eventually found that only 26 of these relations have actual representation in their annotated corpus which are (in order of frequency): *part-whole, attribute-holder, possession, theme, measure, agent, temporal, location/space, kinship, source, topic, recipient, purpose, depiction-depicted, is-a (hypernymy), make/produce, associated with, result, instrument, cause,*

manner, experiencer, means, influence, frequency, and predicate. Later, Girju et al. (2005) and Girju (2007) developed slightly modified versions of this scheme.

Tratz and Hovy (2010) developed their own fine-grained taxonomy of 43 semantic types created mainly by breaking down main semantic categories into subcategories. For example the category of *topic* is divided into 7 other subcategories including *topic of communication*, e.g. "travel story"; *topic of rules*, e.g. "loan terms"; *topic of emotion*, e.g. "jazz fan". They mapped their taxonomy to six previous taxonomies and concluded that the relations are 'fairly similar'. They used five notations for mapping: ≈-approximately equivalent; ⊃/⊂-super/sub set; ∞-some overlap; ∪-union. The problem with this scheme is the obvious high granularity, which could lead to difficulty in making precise distinctions between types.

This overview of the literature on the semantic classification of compound nouns shows the divergence between theoretical and empirical studies. Theoretical studies tend to use coarse-grained types, as they are inclined more towards generalization, while empirical studies tend to use medium- to fine-grained sets of labels to be more descriptive and more exhaustive of the data they target to annotate.

5.2. Our Annotation Scheme

In our work on the semantic annotation of compound nouns (specifically Arabic Idafa), we adopt the 26 taxonomy types designated by Moldovan et al (2004). The reasons we adopt this annotation scheme are: 1) its medium granularity which makes a good compromise between generalizability and applicability; 2) it was designed for English noun phrases which are closer to Arabic IC than the more restricted set of noun-noun compounds; and 3) they use purely semantic notation, i.e. without relying on grammatical functions or lexical interpretation.

Lauer's list of 8 prepositional paraphrases for noun compound has gained popularity in the literature the field since its inception. We argue against the use of prepositions for the purpose of semantic classification of compound nouns, as they are, to a great extent, language specific. For example, the English preposition 'of' does not have a translation into Arabic, 'in', 'at' and 'on' are translated by a single preposition in Arabic 'في' fiy "in". Moreover, the English prepositions 'in', 'at' and 'on' can refer to the same semantic relation of time, e.g. "in the morning", "at night", and "on Sunday". The three prepositions can also indicate the location semantic relation, "in the city", "at the airport", and "on the ground". Therefore, we argue for "conceptual paraphrasing", rather than prepositional paraphrasing.

Another argument against the use of prepositions for denoting semantic relations is that in our framework, lexicon, syntax and semantics are considered autonomous levels of representation/analysis of the language. Therefore, the use of lexical items for the typification of the more universal semantic types should be avoided. Moreover, prepositions are among the class of function words which are poor encoders of semantic content.

The classification of semantic relations in compound nouns can also be conditioned by genre. For example, the following relations can (somewhat) be specific to technical domains: *material*, e.g. "copper wire"; *operation*, e.g. "friction brake" and *representation*, e.g. "regulation curve" (Linh, 2010). In contrast, we are interested in the classification of semantic relations of compound nouns in general domains.

Semantic Relation Type	*%*	*Example*
THEME	28.46	فتحُ الأسواق fatohu Al->asowaqi "opening the markets"
LOCATION/SPACE	8.38	مُحادثاتُ القِمَّةِ muHadavAtu Al-qim~api "summit talks"
AGENT	7.57	إشرافُ اللجنةِ <i$orafu Al-lajonati "supervision by the committee"
PART-WHOLE	6.76	قاعةُ الكنيسةِ qAEapu Al-kaniysapi "the church hall"
TOPIC	5.54	قصائدُ الحبِّ qaSa}idu Al-Hub~i "love poems"
POSSESSION	5.34	مجدُ لبنانَ majodu lubnAna "Lebanon's Glory"
MANNER	5.23	قُصارَى جُهْدِ quSAraY juhodi "best efforts"
PURPOSE	4.52	بطاقاتُ تهنئةٍ biTAqAtu tahoni}apK "congratulation cards"
TEMPORAL	4.12	صَلاَةُ المَساءِ SalAapu AlmasA'i "evening prayer"
ATTRIBUTE-HOLDER	3.76	رائحةُ البرتقال rA}iHapu Al-burotuqAli "smell of oranges"
MEASURE	3.00	نصفُ المدةِ niSofu Al-mud~api "half the period"
CAUSE	2.95	أخطارُ التدخين >axoTAru Al-tadoxiyni "dangers of smoking"

ASSOCIATED WITH	2.74	مُفوَّضُ الحكومةِ mufaw~aDu Al-Hukuwmapi "government commissioner"
IS-A (HYPERNYMY)	1.78	مدينةُ الإسكندريةِ madiynapu Al-<isokandodariy~api "city of Alexandria"
RECIPIENT	1.37	تَدْرِيبَ الإِخْتِصاصِيِّينَ tadoriyba AlAixotiSASiy~iyna "specialist training"
SOURCE	1.37	زَيْتِ الزَيْتُون zayoti Alzayotuwni "olive oil"
EXPERIENCER	1.17	مَخاوِفَ المَسِيحِيِّينَ maxAwifa AlmasiyHiyi~yna "christians fears"
RESULT	1.12	إفْراغ الباخِرَةِ <iforAgi AlbAxirapi "ship Unloading"
DEPICTION-DEPICTED	1.02	واقِعِ الحال wAqiEi AlHAli "the reality of the situation"
MAKE/PRODUCE	1.02	مَنْحُوتَةُ بُول manoHuwtapu buwl "Paul's sculpture"
KINSHIP	0.76	أهالي الشُهداء >ahAli Al-$uhada}i "families of the martyrs"
ACCOMPANIMENT	0.61	مقابلةُ الرئيسِ muqAbalapu Al-ra}iysi "meeting (by) the president"
EXTENT	0.61	صَوْبَ غابَةِ Sawoba gAbapi "towards the jungle"
INSTRUMENT	0.36	مَعاصِرَ الزَيْتُون maEASira Alzayotuwni "Olive presses"
ENTAIL	0.30	مَفاعِيلِ المَرْسُوم mafAEiyli Almarosuwmi "effects of the decree"
OTHERS	0.15	دافِعِ الحِرْصِ dAfiEi AlHiroSi "diligence motivation"

Table 3: 26 TIC Semantic Types and their percentage presence in a random sample of 1946 ATB TIC data

Accordingly, Table 3 illustrates results of our pilot semantic annotation of TICs in Arabic applied to a subset of 1,946 randomly selected examples from the TIC ATB cases. The semantic types are listed in order of frequency in the data.

6. Conclusion

In this paper we have presented an exposition of the Idafa construction, a ubiquitous genitive construction that is used primarily for the formation of noun compounds in Arabic. We develop a tool for extracting and tagging ICs with its 10 syntactic types corresponding to three top level ICs in Arabic, based on morpho-syntactic cues. This classification shows the adjectival and grammatical functions this construction can serve in addition to the typical function of integrating two nouns as a single unit. Furthermore, we review a variety of semantic annotation schemes for compound nouns, select a framework that best suits the language and the task at hand, and apply it in the classification of TIC (true Idafa) in Arabic.

7. Acknowledgment

This work has been partially funded by the DARPA DEFT project with subcontract to GWU. The views and opinions presented here are those of the authors not reflecting those of DARPA. We further acknowledge the valuable reviews rendered by the LREC paper review committee.

8. References

Asghari, H., Maleki, J., & Faili, H. (2014, April). A Probabilistic Approach to Persian Ezafe Recognition. In *EACL* (pp. 138-142).

Barker, K., & Szpakowicz, S. (1998, August). Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1* (pp. 96-102). Association for Computational Linguistics.

Bögel, T., & Butt, M. (2013). Possessive clitics and ezafe in Urdu. *Morphosyntactic Categories and the Expression of Possession*, *199*(291), 86-129.

Boujelben, I., Mesfar, S., & Hamadou, A. B. (2010). Arabic Compound Nouns processing: inflection and tokenization. *In Proceedings of Nooj International Conference* 2010, 40.

Downing, P. (1977). On the creation and use of English compound nouns.*Language*, 810-842.

Gabbard, R., & Kulick, S. (2008, June). Construct state modification in the Arabic treebank. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (pp. 209-212). Association for Computational Linguistics.

Gadalla, H. A., & Abdel-Hamid, A. M. E. (2000). Genitive Constructions in English and Arabic: A Contrastive Study. *In the Bulletin of the Faculty of Arts, Assiut University, Egypt. Vol. 6, Dec. 2000*, pp. 1-64.

Girju, R., Moldovan, D., Tatu, M., & Antohe, D. (2005). On the semantics of noun compounds. *Computer speech & language*, *19*(4), 479-496.

Girju, R. (2007, June). Improving the interpretation of noun phrases with cross-linguistic information. *In the Proceedings of the 45th Annual Meeting of the Association for Computational Linguistic* (Vol. 45, No. 1, p. 568).

Green, S., & Manning, C. D. (2010, August). Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 394-402). Association for Computational Linguistics.

Habash, N., Gabbard, R., Rambow, O., Kulick, S., & Marcus, M. P. (2007, June). Determining Case in Arabic: Learning Complex Linguistic Behavior Requires Complex Linguistic Features. In *EMNLP-CoNLL* (pp. 1084-1092).

Habash, N., & Roth, R. M. (2009, August). Catib: The columbia arabic treebank. In *Proceedings of the ACL-IJCNLP 2009 conference short papers* (pp. 221-224). Association for Computational Linguistics.

Lauer, M. (1995). Designing statistical language learners: experiments on noun compounds. Macquarie University, Australia.

Lees, R. B. (1960). The Grammar of English Nominalizations. The Hague-Paris, Mouton de Gruyter.

Levi, J. N. (1978) The Syntax and Semantics of Complex Nominals. New York: Academic Press.

Lotfi, A. R. (2014). Persian Ezafe as a 'figure'marker: A unified analysis.*California Linguistic Notes*, *39*, 1.

Linh, N. M. (2010). Noun-noun Combinations in Technical English. A Master's Thesis, Suranaree University of Technology, Thailand

Megerdoomian, K. (2000). A computational analysis of the Persian noun phrase. *Memoranda in Computer and Cognitive Science MCCS-00-321, Computing Research Lab, New Mexico State University*.

Maamouri, M., Bies, A., & Kulick, S. (2008, May). Enhancing the Arabic Treebank: a Collaborative Effort toward New Annotation Guidelines. In *LREC*.

Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., & Girju, R. (2004, May). Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics* (pp. 60-67). Association for Computational Linguistics.

Samvelian, P. (2007). The Ezafe as a head-marking inflectional affix: Evidence from Persian and Kurmanji Kurdish. *Aspects of Iranian Linguistics: Papers in Honor of Mohammad Reza Bateni*, 339-361.

Shilon, R., Habash, N., Lavie, A., & Wintner, S. (2012). Machine translation between Hebrew and Arabic. *Machine translation*, *26*(1-2), 177-195.

Strunk, J. (2003). The structure of the Kurdish noun phrase. *unpublished ms*.

Tratz, S., & Hovy, E. (2010, July). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 678-687). Association for Computational Linguistics.

Vanderwende, L. (1994, August). Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th conference on Computational linguistics-Volume 2* (pp. 782-788). Association for Computational Linguistics.

Warren, B. (1978). Semantic patterns of noun-noun compounds. *Acta Universitatis Gothoburgensis. Gothenburg Studies in English Goteborg*, *41*, 1-266.

9. Language Resource References

Maamouri, M., Bies, A., Kulick, S., Krouna, S., Gaddeche, F. & Zaghouani, W. (2010). Arabic treebank: Part 3 v 3.2. *Distributed by the Linguistic Data Consortium.* LDC Catalog No.: LDC2010T08.