AttImam: A Corpus of Arabic Attribution

Amal Alsaif, Tasniem Alyahya, Madawi Alotaibi, Huda almuzaini, Abeer Alqahtani
Al-Imam Mohammad Ibn Saud Islamic University
College of Computer Sciences and Information
E-mails: hamozeani@imamu.edu.sa , aakalqahtani@sm.imamu.edu.sa

Abstract
Language modeling and its applications are influenced by data quality and the richness of the features that are able to extract, especially for low-resource languages such as Arabic. This paper presents the first empirical effort in annotating attribution in Modern Standard Arabic (MSA). We based on the Arabic Treebank (ATB) which is a corpus that constitutes the gold standard annotation for morphological, syntactic, and discourse features as in the LADTB. The identification of attributed arguments to a source, or so called author identification, has been applied successfully in diverse systems such as information retrieval, and opinion/sentiment mining for several languages. Although lexical and syntactic features play a significant role in these applications, there is a huge demand for Arabic attribution resources in long and short text. We, therefore, insight of prior efforts for other languages, annotate several features for Arabic author identification systems: the basic elements of attribution – cue, source, and content spans – in addition to their supplemental information and all related semantic features. Inter-annotator agreement tests were conducted to ensure a reliable gold standard attribution corpus for MSA. Our guidelines and the proposed annotation tool for annotating attribution in MSA might be used for other similar languages with minimal adaptation. The corpus distribution and the findings of this empirical work will certainly improve linguistic studies in Arabic.

Keywords: Author identification, Attribution, Annotation tool, Natural language processing, Arabic discourse, Annotation guidelines, Arabic treebank, Inter-annotator agreement.

1. Introduction

Attribution in computational linguistics refers to the process of reporting or assigning a spoken or written utterance or piece of content to the correct speaker or agent. Regardless of the truth of the utterance or content, identifying its source, the content boundaries, and other supplemental information (e.g., temporal circumstances) plays a key role in many artificial intelligence applications (Prasad et al. 2006; Pustejovsky and Stubbs 2012). Nevertheless, annotated attribution corpora are essential tools for research in many areas of computational linguistics and natural language processing (NLP), especially for supervised machine learning and deep learning approaches.

For example, in authorship identification, sentiment analysis, story generation, question answering, summarization, and opinion mining (Al-Sarem et al. 2020; Guzmán-Cabrera et al. 2009; Juola and

Baayen 2005; Neumann and Schnurrenberger 2009; Ombabi, Ouarda, and Alimi 2020), annotated attributions have been used to develop and evaluate systems. Given the substantial (and rapidly growing) volume of text data available online, automatic annotation of attribution is indispensable, but it means that it is necessary to address various challenging aspects of language. Several projects on discourse structure have examined attribution as a form of discourse relations, including in the English language (Pareti 2012; Prasad et al. 2006, 2007). However, relatively few studies have focused on attribution (Al-Saif et al. 2018; Pareti 2011, 2016), particularly the issue of annotating all related features. Despite these efforts, there is no annotation for Arabic attribution nor an annotated corpus that can drive forward automatic systems related to attribution in Arabic.

Attribution can be presented directly, often using quotation marks (as in Ex.1), or indirectly, without the use of quotation marks (as in Ex.2). Several speech acts, including (said/qAl/قال) and (expressed/AbdaY/أبدى), have additional semantic information on the top of only reporting the speech. These meanings might be expressed lexically or pragmatically by the speech act itself, the reported speech or any additional information conveyed by the writer/speaker. The annotation of attribution is defined by three basic elements: source, cue, and content (linguistic clauses). Attribution may include other features such as polarity and factuality, supplements, source attitude, and determinacy (Pareti 2016). Empirical studies indicate that there are various syntactic structures of attribution not documented in linguistic resources (Van Son et al. 2016).

Ex.1

قال طلاس "الرئيس بشار لا يمكن ان يتنازل عن شيء استمسك به والده طوال حياته"

Tlass **said,** "*President Bashar could not give up something that his father had stuck to throughout his life*."

Ex.2

أبدى مسؤول اهتمامه بضم مهاجم منتخب البرتغال ونادي بنفيكا نونو غوميش (24 عاما)

An official has **expressed** *interest in signing the Portuguese national striker and Benficano Gomez* (24 years old).

This paper presents an empirical annotation study for Arabic attribution along with a gold standard attribution corpus, including the annotation guidelines, annotation tool, and agreement study. The rest of this paper is organized as follows: Section 2 reviews annotation-related works; Section 3 highlights studies that have focused on attribution in Arabic literature, as well as related language modeling studies; Section 4 introduces the proposed annotation schema of attribution, along with reasonable examples; Section 5 presents the annotation process, the tool, and the inter-annotator agreement studies for all elements; Section 6 discusses corpus distribution and disagreement cases; and finally, Section 7 concludes the paper with a discussion of future applications and recommendations for further research.

2. Related Work

Several discourse theories have dealt with attribution relations in different phenomena and point view as that attribution might lead to a mismatch between the syntactic and discourse arguments of discourse connectives (Dinesh et al. 2005). An overview of relevant resources annotated with certain types of attributions or other relations overlapping with attribution is presented in (Pareti 2016). For instance, the Rhetorical Structure Theory (RST) Discourse Treebank (Carlson and Marcu 2001) has few attribution relations in 385 *Wall Street Journal* (WSJ) articles using the RST framework (Mann and Thompson 1988). In addition, Graphbank (Wolf and Gibson 2005) has 1,400 attribution relations in 135 texts from AP Newswire and WSJ. Both discourse studies encoded attribution relations by annotating two elements: (1) the attributed span, and (2) the attribution span (i.e., John said, [attributed span] "You are so smart" [attribution span]). However, the first study annotated only intra-sentential attributions with an explicit source and a verb cue, and the latter annotated attributions if no other discourse relation was present.

The comparison in (Pareti 2016) indicates that there are two large resources that comprise attribution annotations: the Penn Discourse Treebank (PDTB) (12,000 attributions in 2,159 news articles) and opinion corpus: Multi-Perspective Question Answering (MPQA) in 535 news texts (Wiebe 2002); and the Columbia Quoted Speech Attribution (CQSA) (Elson and McKeown 2010), which annotates 3,500 opinions in 11 narrative books. The opinion annotation comprises beliefs and opinions, as well as the quotations introducing them, but not the content span boundaries and other detailed features of attribution. By contrast, in the PDTB, attribution is annotated along with explicit and implicit discourse connectives and their arguments as separated relations or as discourse relations. However, when part of the attributed span does not correspond to the discourse argument of the relation, annotated attributions are missing or incomplete. Also, the proposed schema overlooked nested attributions and did not specify the purpose of the attribution, which is essentially inferred by the cue and other supplemental information. Additional research efforts that have addressed elements of attribution annotation are discussed in (Pareti 2016).

(Pareti 2016) proposed a solid basis for a new schema for attribution in three projects for the Penn Attribution Relation Corpus (PARC) (Almeida, Almeida, and Martins 2014; O'Keefe et al. 2012; Pareti 2016; Pareti et al. 2013), the last of which was PARC 3.0, generating by using the PDTB as a ground definition. Similar to other projects, three basic components of attribution were annotated (i.e., the cue: text anchor, the source: agent owning the content, and the content: the quoted text span), along with an additional fourth component: the supplement, which denotes any extra information expressing the attribution (e.g., circumstantial information). PARC 3.0 also annotates features included in the PDTB such as attribution type, source type, factuality, and scopal polarity. Most attribution corpora, with the exception of the Italian Attribution Corpus (ItAC) (Pareti and Prodanof 2010), which annotates a pilot corpus of 50 articles, are for the English language.

Attribution was also a part of a corpus of perspectives in a text in the study of (Van Son et al. 2016), which used a model known as Grounded Representation and Source Perspective (GRaSP). This multi-layered annotation approach has four different elements: events, attribution, factuality, and opinion. The attribution is addressed in the second layer by defining the three main elements of a factuality relation:

source, cue, and target. The researchers followed the definition of attribution relations provided in the PDTB and the PARC projects.

Empirically, authorship attribution is a subfield of automatic authorship analysis that associates an anonymous text with an author based on certain features. Several studies have examined automatic quotation on narrative texts, including (Mamede and Chaleira 2004) and (Elson and McKeown 2010), and others have addressed the topic on news texts (Pouliquen, Steinberger, and Best 2007; Sarmento, Nunes, and Oliveira 2009). These studies were based on the use of lexical and syntactic rules to infer the author of the quoted text. Experimental evidence in these studies clearly indicates the partial usefulness and, therefore, the unreliability of the proposed systems across various domains and writing styles. The later research by (Elson and McKeown 2010) and (Fernandes, Motta, and Milidiú 2011) was more successful because the researchers exploited machine learning models with relatively large, annotated corpora. In (Elson and McKeown 2010), the researchers used a corpus of approximately 3,000 quotes and manually identified speaker candidates. The proposed approach achieved an average accuracy of 83% using lexical and syntactic features. In (Fernandes, Motta, and Milidiú 2011), the corpus they created contained annotated golden features for entities, coreferences, quotes, associations between quotations and authors, and Part-of-Speech (POS) tagging information. For the Portuguese language, the accuracy of author attribution was 79%.

Machine learning approaches for identifying authors in long texts typically use two types of features: firstly, Bag-of-Words (BOW) features, where the feature vector for each document is computed based on word frequencies; and secondly, Stylometric Features (SF), which allow differences between the writing styles of different authors to be captured and inferred. The writing styles of each author can be further categorized based on their lexical, structural, syntactic, and content-specific features (Altheneyan and Menai 2014; Cheng, Chandramouli, and Subbalakshmi 2011; Koppel, Schler, and Argamon 2009; Pillay and Solorio 2010; Tan and Tsai 2010). The work by (O'Keefe et al. 2012) overcame several unrealistic gold standard features in previous studies by using three sequence decoding models, including greedy, Viterbi, and a linear chain Conditional Random Field (CRF) on two news corpora: the WSJ and the Sydney Morning Herald (SMH). The highest accuracy rates were 84.1% and 91.2% for WSJ and SMH, respectively.

3. Attribution in Arabic

Modern Standard Arabic (MSA) was developed in the late nineteenth century, and it is used by non-Arabic linguists to refer to the Arabic literary (Badawi, Carter, and Gully 2013). Today, MSA is the official language for Arab countries, which is used in academia, government, law, legislation, media, and among others. However, MSA is not the first language that many residents of Arab countries acquire (Habash 2010). The justification for MSA is that it is intended as a flexible version of Classical Arabic (CA), which is the language of the Quran and classical Arabic literature. Therefore, MSA can be considered a simplified version of CA that seeks to retain the same basic syntax and morphology while coping with the demands of modern society. Today, Arabic natural language processing (NLP) research focuses on MSA, and many language-specific tools and resources have been developed for different areas

such as machine translation, information analysis, and retrieval (Al-Sarem et al. 2020; Alam and Kumar 2013; Otoom et al. 2014; Rabab'ah et al. 2016).

There is a base of attribution in the traditional literature on semantic science (semantics/Elm AlmEAny/علم المعاني), and speech reporting (report/AsnAd/اسناد-report/nql klAm/نقل الكلام-report /rwAyp/رواية) is an essential area in writing styles. In (Ywns 2004), the basic structure of speech reporting is studied mainly through the use of lexical, syntactic, or pragmatic features that are not connected to any discourse theories. As in many languages, reported speech in Arabic can occur in the form of direct paraphrasing (i.e., the exact words of the reported speech using quotations) and indirect paraphrasing (i.e., the semantic content of the reported speech), where it can appear in up to 90% of the latter (Al-Saif et al. 2018; Pareti 2016). Direct reporting is identified by the existence of typographic indicators (e.g., colon and quotation marks), whereas indirect reporting is distinguished through a combination of linguistic indicators without any typographic ones (e.g., verbs and adverbials). An example of two different ways to introduce reporting speech is worth considering: firstly, using quotation marks, as in (He said: "The book is interesting"/qAl: "AlktAb m$wq" /"قال: "الكتاب مشوق"); and secondly, using only linguistic markers as the combination of the words (said/qal/قال) and (that/An/إن) in (He said that the book is interesting/qAl An AlktAb m$wq/قال إن الكتاب مشوق). Indirect speech reporting is clearly more complex than direct speech, not only in the form of reporting but also in the interoperation of the reporter's intention when reporting the speech. The indirect speech may lead to pragmatic usage of the reporting, which may result in misunderstanding. This highlights the challenge facing automatic identification for attribution comments.

(Speech Act Theory/nZryp >fEAl AlklAm/نظرية أفعال الكلام) (Searle 1976) builds on ideas proposed by (Austin 1975). It is a reputable pragmatic concept that has aided in identifying and understanding the semantics of speech reporting. MSA used this theory to gain a thorough understanding of the semantics of reporting speech and related issues (i.e., reporting elements and the relationship between the source and writing style). One of the main contributions of (Searle 1976) is the discrimination between two types of speech acts: firstly, constative verbs; and secondly, performative verbs. The former are propositional sentences that can be stated to the truth value (e.g., she said: "The trees are green"/qAlt: "Al>$jAr xDrA'"/"قالت: "الأشجار خضراء"). The sentences present a statement that can be assessed based on the truth at that time with no action by the source. By contrast, performative verbs refer to those sentences that express performing an action. Performative verbs can be assertive (e.g., suggest/AqtrH/اقترح), directive (e.g., ask/As>l/اسأل), commissive (e.g., promise/wEd/وعد), expressive (e.g., forgive/AEtdr/اعتذر), or declarative (e.g., confirm/Akd/أكد). The MSA literature showcases the tremendous efforts that have been made to distinguish between constative and performative sentences (Matloob 1986). Moreover, attempts have been made to identify the most frequent Arabic speech acts (BwEyAd and Blxyr 2012). It was not until 2018, with the publication of (Al-Saif et al. 2018) paper that closely followed (Pareti 2016) approach, that the basic grounds were established for defining MSA attribution for direct and indirect speech and clearly identifying its elements. This paper continues reporting all adaptations and results by empirical annotation in order to construct a gold standard attribution corpus for Arabic.

In the domain of computational linguistics, existing Arabic NLP systems do not usually target reported speech itself. They mainly focus on automated retrieval of quoted segments without going further into

either the enunciative modality or the linguistic analysis of the introduction markers of the reported speech. One reason for this is the absence of an annotated corpus for attribution in Arabic, which covers both direct and indirect speech with their features. Most studies have used datasets consisting of long pieces of domain-specific text or short texts (e.g., Twitter posts).

For short texts consisting of tweets written in Arabic, (Rabab'ah et al. 2016) examined author identification in the context of social media. The dataset consisted of 38,386 tweets for 12 users. The researchers used Naive Bayes (NB) and Support Vector Machine (SVM) based on 340 stylometric features (SF) and 57 morphological features (MF) in mostly POS-based features. In the experiment, SVM showed the highest accuracy on the combined feature sets, which was 68.67%. In a later study, (Al-Ayyoub, Alwajeeh, and Hmeidi 2017) extended the experiment by applying the BOW approach to derive a third feature set. They derived the same result for the SVM classifier in terms of accuracy, and no significant advantage was identified for SubEval features with the exception of a reduction in the classifier in building time by 93%.

For long texts, (Alam and Kumar 2013) addressed the authorship attribution problem by proposing a novel Arabic feature extraction system that considered semantic features from morphological structure. The dataset consisted of 100 articles evenly distributed across 10 authors. The highest accuracy was 98%, which was achieved by applying the semantic features and SVM_light algorithm. In (Otoom et al. 2014), other experiments were conducted with five classifiers, including BayesNet, MultiBoostAB, NB Random Forest, and SVM. A hybrid set of four types of features was extracted from the collected Arabic text articles: lexical, syntactic, structural, and content-specific features. The dataset consisted of 456 articles for 7 Arabic writers, and it covered topics ranging from politics to sport and others. The experiments were conducted with the hold-out test, and an accuracy of 88% was achieved with the MultiBoostAB method, thereby proving the robustness of the proposed feature set for long texts. The authorship authentication problem was addressed by (Alwajeeh, Al-Ayyoub, and Hmeidi 2014). The dataset consisted of 500 articles distributed over 5 authors. The highest accuracy was 99.8%, achieved using SVM, whereas for NB, the accuracy was 99.4%.

Recently, a study was undertaken that involved combining a set of classifiers and voting their prediction for authorship identification (Al-Sarem et al. 2020). Two types of features were used: firstly, stylometric features; and secondly, distinct words from a large fatwa corpus, which was assembled from the *Dar Al-ifta AL Misriyyah* website. The results indicated that the AdaBoost methods obtained the highest accuracy for the balanced dataset, whereas the Bagging methods obtained the highest accuracy with the unbalanced dataset. The authors also offered a review of recent initiatives using the manual collection of labeled articles with the author and their style in Arabic articles, including (Ouamour and Sayoud 2018) and (Ahmed, Mohamed, and Mostafa 2019). Further, they emphasized the importance of an attribution corpus with high-quality annotation of morphological, syntactic, and semantic analysis.

In the same context, (El Bakly, Darwish, and Hefny 2020) used an Islamic fatwa corpus for "traveler's prayers" in the main Islamic jurisprudence doctrines of Hanfi, Malki, Shafie, and Hanbali. The dataset consisted of 1,073 fatwas, which were partitioned into two groups: firstly, 751 fatwas for a training set; and secondly, 322 fatwas for a test set. The aim of the study was to attribute the unknown fatwas to one

of the main Islamic jurisprudence doctrines and to create a new corpus of the traveler's prayer fatwas. In their work, they applied a novel model based on employed ontology as a semantic feature. In their experiments the achieved performance for their proposed method was 90%.

4. Attribution Schema of Arabic

The proposed annotation schema for attribution relation in Modern Standard Arabic (MSA) is presented in this section. Our guidelines are based on the work of the Penn Discourse Treebank (PDTB) (Prasad et al. 2006) and the Penn Attribution Relation Corpus (PARC) (Pareti 2011, 2012, 2016; Pareti and Prodanof 2010), with certain adaptations and extensions to suit MSA. The definition of attribution relations includes three basic elements: cue, source, and content. In addition, reporting an event or discussing a topic usually involves other individuals' speech with supporting materials, including temporal cues and locations, which lead to additional elements (i.e., supplemental information). These supporting materials are usually used to provide a detailed description of entities or events, or to support facts that are mentioned in the content. Each basic element may have related supplemental information. For example, the temporal text expresses the timing of the attributed cue.

The four constitutive classes are the cue, the source, the content, and the general features of the attribution. It is worth mentioning that an attribution relation is an intra-sentential relation, which is a mixture of syntactic and semantic relations between words and clauses in the sentence. The annotation manual presents each class with a clear definition and examples using the following convention: the cue is bold-faced, the source span is underlined, and the content span is italicized. We differentiate between the types of supplemental information as follows: the cue supplement (enclosed in brackets); the source supplement (enclosed in braces); and the content supplement (enclosed in parentheses).

4.1 Cue

The **cue** can be defined as the lexical anchor that connects the source with the content. Various syntactic forms of the cue have been distinguished in literature and from our pilot annotation as follows (Appendix A lists all cue types with examples):

- A reporting verb/speech act that is either constative or performative, such as (said/qAl/قال) in Example 1; (stated/ SrH/صرح) in Example 2; and (confirmed/ >kd/أكد) in Example 3.
- Adverb works as a reporting verb, such as (declaring /mElnA/معلنا) in Example 6 and (explaining/mwDHA/ موضحا) in Example 7.
- A prepositional phrase works as implicit reporting verb, such as (according to/bHsb /بحسب) in Example 8.
- Absent cue (or implicit cue), such as (Sara: I will study today/ sArh: swf >drs Alywm / سارة: سوف أدرس اليوم), which is understood from the context as (*said/* qAl /قال).

Ex.3
قال ديميتري لاشوتين {لمسؤول عن المركز الحكومي لتنمية السياحة في التاي} [لوكالة فرانس برس] *"عام 1999 زارنا نحو 300 الف سائح غالبيتهم من الروس الذين قدموا من مناطق مجاورة -كيميروفو ونوفوسيبيرسك- ونحو الف اجنبي فقط الا اننا نامل بان يزيد هذا الرقم سريعا"*
Dmitry Lashotin {the official of the State Centre for Tourism Development in Tai} **said** [to the France Press Agency]: "*In 1999, we visited about 300,000 tourists, most of them Russians who came from neighboring regions – Yerovo and Novosibirsk – and only about a thousand foreigners. This number increased rapidly".*
Cue Status: Explicit, Cue Purpose: Assertion, Attribution Style: Direct, Determinacy: Yes
From: 20000715_AFP_ARB_0030.raw

Ex.4
صرح الوزير الإيراني [للتلفزيون] *"لقد اعتبرنا ان اجتماع اوبك سيكون سابقا لأوانه لأننا يجب ان ننتظر ونرى تطور السوق".*
The Iranian minister **stated** [to television]: "*We have considered that the OPEC meeting will be premature because we have to wait and see the development of the market".*
Cue Status: Explicit, Cue Purpose: Declaration, Attribution Style: Direct, Determinacy: Yes
From: 20000715_AFP_ARB_0046.raw

Ex.5
أكد الرئيس العراقي صدام حسين *حرص العراق على تطوير العلاقات مع الجارة إيران* (حسبما ذكر الوزير الإيراني).
Iraqi President Saddam Hussein **confirmed** *Iraq's keenness to develop relations with neighboring Iran* (according to the Iranian minister).
Cue Status: Explicit, Cue Purpose: Assertion, Attribution Style: Indirect, Determinacy: Yes
From: 20001015_AFP_ARB_0100.raw

Ex.6
معلنا *ان انفجار المدمرة الاميركية "عمل رهابي".*
Declaring *that the explosion of the American destroyer is a "terrorist act".*

Cue Status: Explicit, Cue Purpose: Assertion, Attribution Style: Indirect, Determinacy: Yes
From: 20001015_AFP_ARB_0164.raw

Ex.7
موضحا انه تلقى تعليمات بذلك من الخطوط الجوية السعودية.
Explaining that he *had received instructions to do so from Saudi Airlines.*
Cue Status: Explicit, Cue Purpose: Assertion, Attribution Style: Indirect, Determinacy: Yes
From: 20001015_AFP_ARB_0017.raw

Ex.8
ولم يتم اعتقال أحد **بحسب** المتحدث باسم الشرطة.
No one was arrested, ***according to*** *the police spokesman*.
Cue Status: Implicit, Cue Purpose: Declaration, Attribution Style: Indirect, Determinacy: Yes
From: 20000715_AFP_ARB_0076.raw

The **cue status** feature has two options: either *Explicit* to specify whether the cue explicitly occurred in the text, or *Implicit* in case the cue is not mentioned in the sentence. To distinguish between these two types of cue status more effectively, it should be noted that if the attribution has two verbs, the first of which is an attributed span and the other part being the verbal phrase in the content, then it is an *Explicit* cue. By contrast, if there is only one main verb, then the first is the cue, and this can be considered an *Implicit* cue. *Implicit* cues are used often with indirect attribution with no quotation marks, where there is only one main verb in *performative* speech acts, expressing the attribution and the action/fact in the content simultaneously. The cue in such cases will be included in the content, such as *She demands additional investigation /* تطالب بتحقيق اضافي and *he promised more attention /* وعد بمزيد من الاهتمام.

The *Explicit* cue is used with either direct speech, when the content is introduced by punctuation marks (in this case, : or ""), or with indirect speech, when it is not possible to determine whether the content has exact words of actual speech. The indirect speech is commonly associated with particles (that/An/ان). To distinguish the *Implicit* cue from *non*-cue verbs,[1] one suggestion to the annotators is to convert the *Implicit* cue into a direct attribution using a reporting verb (e.g., such as said/qAl/قال), and then to identify the

[1] All verbs are pre-highlighted in the tool.

content boundaries between quotation marks. If the conversion is achieved successfully without changing the meaning of the reporting or generating redundancy, then it is an *Implicit* cue.

Consider the example of the verb "demand" in (*She demands additional investigation /* تطالب بتحقيق اضافي). To determine whether it is an *Implicit* cue, we convert the sentence into direct speech, which produces *(She said: "I demand an additional investigation"/*قالت :"أطالب بتحقيق أضافي*").* If the meaning of the sentence is clear with no redundancy, then it is an *Implicit* cue; otherwise, there is no attribution. Examples 3 and 5 show an *Explicit* cue with direct speech and indirect speech, respectively, while Examples 8 and 9 show an *Implicit* cue with indirect speech.

Ex.9
تطالب *بإشراكها في مفاوضات السلام.*
She ***demands*** *her participation in peace negotiations.*
Cue Status: Implicit, Cue Purpose: Directive, Attribution Style: Indirect, Determinacy: Yes
From: 20000815_AFP_ARB_0078.raw

Although it is not strictly part of the cue, the **cue supplement** is the text span that is relevant to the interpretation of the cue. Usually, the narrator is keen to transfer the circumstantial occurrence of the attributed text in order to confirm the credibility of the transfer, on the one hand, and to address the expectations of the recipient and the knowledge background (thus understanding the attributed content), on the other hand. The forms for the cue supplement are as follows: adverbs expressing a status such as (laughing/DAHkA /ضاحكا) and (in say while laughing/qAl DAHkA / قال ضاحكا); temporal circumstances such as (afternoon/zuhraan /ظهرا) and (evening/ masa'/مساء); or a particular location, such as (in Riyadh/fi alriyad/في الرياض). However, certain modifiers, including prepositions such as (on/ ElY/على) and (states that/ nS ElY /نص على), are tagged as part of the cue if they do not express extra semantics to the attribution. Examples 3 and 4 contain a cue supplement, but Example 5 does not contain supplemental information about the cue.

Cue negation is a feature that can be used to determine whether the cue is modified by one of the negation tools, such as *did not/lm/*لم, or whether the cue itself is a lexical verb indicating negation, such as denied/ rfD/رفض. The significance of the cue in terms of proof and negation plays an important role in determining the significance of the narration of the attribution and constructing a discourse relation with the preceding and attached texts. Additionally, it is used in sentiment analysis of the attribution.

Table 1: Negation in an Arabic context

Negation	With negation tools/أدوات النفي	*No/lA/لا*
		Will not/ Ln/لن
		Not/ lys/ليس
		Did not /lm/لم
		Did not / ma/ما
	Lexical negation verbs/أفعال تفيد النفي	*Denied/ rfD/رفض*
		Gainsay/>nkr / أنكر
		Disclaim /nfY/نفى

Table 1 presents several examples of both types of negation forms in Arabic. In Example 10, the cue is negated using a negation tool (no/lm/لم), whereas in Example 11, the cue is the negation verb semantically.

Ex.10
*لم **تفد** معلومات عن اضرار او ضحايا حتى الان.*
***Did not report** any information about damages or casualties yet.*
Cue Status: Implicit, Cue Purpose: Declaration, Attribution Style: Indirect, Determinacy: No
From: 20000815_AFP_ARB_0037.raw

Ex.11
*<u>السلطات الروسية</u> **رفضت** كل عروض المساعدة الخارجية، الاميركية والبريطانية، (التي تهدف الى انقاذ الغواصة التي تحمل 24 صاروخا تسياريا غير مجهزة برؤوس نووية).*
*<u>The Russian authorities</u> **have refused** all offers of foreign aid, American and British (which aims to rescue the submarine carrying 24 ballistic missiles not equipped with nuclear warheads).*
Cue Status: Implicit, Cue Purpose: Assertion, Attribution Style: Indirect, Determinacy: Yes

From: 20000815_AFP_ARB_0088.raw

The **cue digression** feature occurs when the cue digresses from a formerly reported text, such as in Example 12, where the attribution could not initially stand alone. The cue here is usually not a verb; instead, it is an adverb acting as a verb, such as (adding/mudifaan/مضيفا, pointing/mushiraan/مشير). Correspondingly, the attribution purpose here should share the same properties of the preceding attribution.

Ex.12
أعلن موغابي [للصحافيين] *ان التشكيلة الحكومية الجديدة "تعكس وجهة نظرنا الجديدة وتأخذ في الاعتبار ضرورة تقليص نفقات الحكومة"، مضيفا ان" الانخفاض في الوقت الراهن متواضع ولكن اعادة الهيكلة ستستمر" .*
Mugabe **announced** [to reporters] *that the new cabinet "reflects our new viewpoint and takes into account the need to reduce government expenditures"*, **adding** that *"the decline at the present time is modest, but restructuring will continue".*
Cue Status: Explicit, Cue Purpose: Declaration, Attribution Style: Direct, Determinacy: Yes
From: 20000715_AFP_ARB_0061.raw

4.2 Source

In certain cases, identifying the source of the reported text is not trivial, especially if the writer uses implicit cues in indirect speech. To overcome this ambiguity in the annotation process, we define four types of sources with examples in the manual.

The source is defined as the entity or the agent that owns the content. Syntactically, the source is often the subject of the reporting verb (i.e., the cue, whether *Explicit* or *Implicit*). Similar to the PDTB annotation (Prasad et al. 2006, 2007) and PARC (Pareti 2016), the source is annotated by specifying the source type, as well as a word or phrase that expresses the source. The source type expresses diverse types of agents: (i) the author or writer of the text (WR); (ii) any specific agent other than the writer that explicitly occurs in the text (EXP-AG); (iii) implicit agent when it is referred to from former sentences (IMP-AG); or (iv) the source is annotated as (MISS), in which case the writer did not refer the speech to a specific agent.

Generally, the source is WR when the subject of the speech act, the cue, is the writer. After attributing all the text in an essay, the remaining text is by default attributed to the writer. Additionally, EXP-AG occurs when the source is introduced in the attributed sentence, such as in Examples 3 and 12. An EXP-AG can also be an adjective, as in Example 4. Mostly in digression attribution relations, the source is referred to by preceding sentences, as in Example 12, where the source is attributed as IMP-AG for the cue (adding/mudifaan/مضيفا). The final source type is when the agent is (anonymous/mubni lilmajhul/ مبني للمجهول), as in Example 13; in this case, the source is attributed as MISS.

Ex.13
***يذكر** أن صاحبي المركزين الاول والثاني فقط يتأهلان الى سيدني.*
It **is noteworthy** *that only the first and second place winners qualify for Sydney.*
Cue Status: Explicit, Cue Purpose: Declaration, Attribution Style: Indirect, Determinacy: Yes
From: 20000815_AFP_ARB_0051.raw

The **source supplement** feature tags any additional expression related to the source itself. Any supplemental information specifying the time or place is related to the cue and not to the source. For example, any relative clause describing the agent, as in Example 3, is a source supplement. It is noteworthy that the supplement span should include the description of the source after the source span, in Arabic. If the description appears before the source, such as in the phrase (Prime Mister Ahmad/ رئيس الوزراء أحمد), then it is a part of the source itself and is not considered a supplement.

4.3 Content

The content is the basic element expressing the claim or the reported news, and it occurs in the form of a textual statement attributed to the source. The attributed material might range from one word to multiple sentences. In explicit attribution, the content must be either between quotation marks (“”) or after a colon (:), which is explicitly direct (as in Example 3). Alternatively, the content must start explicitly with the particle “that/ >n /أن” as indirect attribution (as in Examples 7 and 13). By contrast, the boundary of the content in the implicit cue is not clear, and thus, the whole text with the cue and the source is annotated as content in order to preserve the syntactic structure of the sentence, as in Examples 9 and 10.

Content supplement: This relates to any additional clauses that are relevant to the content, but that are not part of the reported speech. The content supplement is added by the writer to provide background information or more justification relating to the entities or events in the content. The content supplement is not frequently used in reporting speech and it does not appear in the middle of the content (usually, it appears at the end of the content). An example of this is the relative clause with the pronoun (which/ Alty /التي), as in Example 11.

Content negation: The content can also be negated using either negation tools, such as (no/ lm /لم) in Example 8 or negation verbs in verbal phrases, as shown in Table 1. Negation nouns, such as (refusal/ AlrfD /الرفض) and (denial/ AnkAr /انكار), in the content itself make this feature true.

4.4 General Features

In addition to textual features (cue, source, content, and their supplements), the attribution relation has other important semantic features related to the attribution in general (i.e., rather than to specific elements). Encoded semantic features are useful in many areas, including information/opinion extraction systems. Therefore, we annotate these features with the attribution to offer a rich resource for different language modeling tasks and their applications. The general features include attribution style, determinacy, and attribution purpose.

The **attribution style** feature, which can either be direct or indirect, determines whether the reported text contains the exact words of the speech (direct), as in Examples 3 and 12, or whether it involves paraphrasing (indirect), as in Examples 5, 9, and 11. Direct speech often presents the exact words with punctuation marks (: or ""), whereas indirect speech does not use quotations but may use the particle (that/ >n /أن). Generally, direct attribution involves using an explicit cue and content, whereas indirect attribution often uses implicit cues (see the description of the cue element).

Table 2: Modified tools for hypothesis and future tense in Arabic

Semantic function	Lexical modified tools
Hypothesis/ أدوات الشرط الغير جازمة	إذا/ < **A/if* لو/*if/ lw* لولا/*lwlA / if-not* لوما/ *if-not/ lwmA*
Future tense/ فعل مضارع يدخل عليه الأتي:	س/*will/s* سوف/*shall/swf* قد/*may/qd* ربما/ *rbmA/ might*

The **determinacy** is a feature compliant with the PDTB guidelines, which determines whether the attribution relation is factual or non-factual, regardless of the truth status of the content. The attribution becomes a non-factual relation (indeterminacy) when the attributed span is in the scope of hypothesis, future tense, or negation. In case of the negation, we must be sure that the negation for the cue itself not for the content, such as Example 10 the cue (did not report /lm tfd/لم تفد) is negated, consequently the attribution relation is annotated as indeterminacy. Table 2 presents the function words as modifiers to the cue in conditional sentences and the future tense. Accordingly, Example 14 presents the attribution relation in the future tense with a future keyword (will/sa/س). In hypothesis reporting speech, the conditional function words, such as (if /<lw/لو), contribute to the indeterminacy of the attribution relation, as in Example 15.

Ex.14
سيطلب [من الرئيس الفلسطيني ياسر عرفات] *وقف العنف والتحريض على العنف واعادة اعتقال ارهابيي حماس.*
He will **ask** [Palestinian President Yasser Arafat] *to stop violence and incitement to violence, and to re-arrest Hamas terrorists.*
Cue Status: Explicit, Cue Purpose: Directive, Attribution Style: Indirect, Determinacy: No
From: 20001015_AFP_ARB_0223.raw

Ex.15
لو **أخبرنا** عمر البشير *أنه سيعبر أجواءنا المضمخة بالروحانية هذه الأيام ,كنا سنحتفي به جوا كما يليق بفخامته.*
If Omar al-Bashir **told** us that *he will pass our atmosphere of spirituality these days*, *we will celebrate as worthy of his leadership.*
Cue Status: Explicit, Cue Purpose: Declaration, Attribution Style: Indirect, Determinacy: No
---

The **attribution purpose** is a feature that signifies the nature of the relation between an agent and the cue. It describes the reason for using this particular cue in the reported speech. Based on well-established linguistic theories, including Speech Act Theory (Searle 1976), we classified speech acts (cues) into five distinct purposes in a flat taxonomy of attribution purposes, which is given as follows:

(i) **Assertion:** The source tends to emphasize their opinion or commitment to the truth of the proposition in the content (e.g., said/qal/قال, assert/Akd/أكد, mention/dkr/ذكر).
(ii) **Directive:** The attribution content presents an order or a request conveyed by the source to others, such as the actions of requesting (e.g., order/Amr/أمر request/dlb/طلب) or questioning (e.g., ask/sAl/سأل , question/ Astfhm /استفهم).
(iii) **Commissive:** The source reports their obligation to do something (e.g., bet/ rAhn /راهن, pledge/ wEd /وعد , and oath/ >qsm /أقسم).
(iv) **Expression:** The source expresses their feeling or regret (e.g., apologies/ AEt*r /اعتذر, congratulate/ hn> / هنأ, thank/ $kr / شكر).
(v) **Declarative:** The source declares something without emphasizing their opinion, feeling, or obligation (e.g., announced/ >Eln /أعلن, informed/ >blg/ أبلغ, stated/ >fAd /أفاد).

Explicit attributions, both direct and indirect, often express Assertion and Declarative speech, such as in Examples 3 and 4. The other attribution purposes (Directive, Expression, and Commissive) are commonly

used for implicit indirect attribution. Our empirical study will highlight all significant usages of speech acts with examples, which represents one of the fundamental contributions of this research.

5. Building the AttImam: Attribution Corpus for Arabic

Corpus

The Arabic Treebank is frequently used in Arabic natural language processing (NLP) studies and applications. It is a large-scale corpus annotated for morphological and syntactic annotation. The Arabic Treebank (ATB) was expanded by the Leeds Arabic Discourse Treebank (LADTB), which was initiated by (Al-Saif and Markert 2010), in order to include explicit discourse relations. Given that it is vital to support research in computational linguistics, we decided to expand again the ATB by annotating the attribution relations in the LADTB, which contains 536 news articles.

Annotation Methodology/Process

Human annotation is the most precise way to generate high-quality, gold standard language resources, particularly for languages such as Arabic that are low-resource languages. Later, this gold-standard resource might be used to generate automatic or semi-automatic identification or retrieval systems. However, annotating a large number of texts is time-intensive and costly. Therefore, the more intelligent the annotation tools available, the greater the utility in facilitating a reliable annotation.

Although there are general-purpose annotation tools such as GATE (Ide and Suderman 2009), BRAT (Stenetorp et al. 2012), MMAX2 (Müller and Strube 2006), and WebAnn (Yimam et al. 2013), certain assistive features (e.g., annotation using radio buttons or dropdown lists, which increases the efficiency and convenience of the process) are not supported in these tools. This is especially the case for right-to-left script languages such as Arabic. The pre-annotation process is also not supported (e.g., in terms of tasks such as highlighting lexical items in a given text, which can greatly assist annotators). Most available tools are restricted to only general tasks (i.e., POS tagging and shallow tokenization). In addition, to the best of the author's knowledge, there is no annotation tool that provides an integrated study of agreement with several measures for label and text span attributes.

For the reasons given above, (Al-Saif et al. 2018) developed the ESNAD system (Extracting Sentence Attribution in Arabic Discourse). This annotation tool was developed for annotating attribution in Arabic. Noteworthily, the tool provides valuable annotation features and, furthermore, can compute all types of inter-annotator agreement measures (e.g., exact match, accuracy, F-score (Joshi 2016), Kappa (Siegel 1956), and Agr (Artstein and Poesio 2008) for text attributes) by only selecting the files of the two annotators, as in Figure 2. The main interface of ESNAD is shown in Figure 1.

Annotation Process

Two native Arabic speakers with a background in linguistics performed the annotation on 537 files from the LADTB, which is a subset of ATB Part 1. As mentioned before, this corpus will contain all layers of text analysis after annotating attribution. After applying all adaptations to the annotation guidelines and

the tool as a result of the pilot annotation in (Al-Saif et al. 2018), the annotator uploads the desired file. In turn, all potential cues (which are essentially verbs) are highlighted,[2] and they are listed in order to assist in making a decision about whether each one is a correctly attributed span (*IsCueAttribution*, see Figure 1). Then, the annotator completes all related attributes for the attribution span using the instructions provided in the annotation manual. The manual can also be reached via the help menu in ESNAD. For any additional information or comments, the annotator can write notes related to the attribution. The text fields should be filled only by marking the desired text from the text area and then clicking on the arrow button. The tool will automatically copy the text and save the indices of the boundaries.

The cue (announced/ >Eln/أعلن) in Example 16 is an attributed span, and the supplement is during the journalism conference/xlAl m&tmr SHfy/ خلال مؤتمر صحفي. The annotator highlights (Foreign Minister/wzyr AlxArjyp/وزير الخارجية) as the source of the attribution, with no supplement related to the source. According to our guidelines, the source type is EXP_AG because the source is explicitly mentioned. Likewise, the phrase (*that the talks he had held in Baghdad resulted in an agreement to revive the old committees and resolve all related issues/* ان المحادثات التي اجراها في بغداد اسفرت عن الاتفاق عن احياء اللجان القديمة وحل جميع القضايا المتعلقة) is marked with no supplemental information too. In addition, there is no negation of the annotated span and content.

Ex.16
أعلن وزير الخارجية [خلال مؤتمر صحفي] *ان المحادثات التي اجراها في بغداد اسفرت عن الاتفاق عن احياء اللجان القديمة وحل جميع القضايا المتعلقة*
The Foreign Minister [during the journalism conference] **announced** *that the talks he had held in Baghdad resulted in an agreement to revive the old committees and resolve all related issues.*
Cue Status: Explicit, Cue Purpose: Declaration, Attribution Style: Direct, Determinacy: Yes
From: 20001015_AFP_ARB_0068.raw

In the last panel, the annotator must annotate three general attributes: attribution style, cue determinacy, and attribution purpose. In Example 16, the style of attribution is indirect because it does not represent the exact words of the source. In addition, the determinacy is "yes" as the attribution is factual, while the purpose of attribution is declaration because there is no emphasizing, feeling, and the source simply reports the speech.

[2] 2,600 verbs were extracted from ATB.

It is essential to note that the annotator is not restricted to the suggested verbs as potential cues. Instead, the annotator can annotate any text span as a cue by selecting the desired cue and clicking on the “New Cue” button, and then completing the annotation.

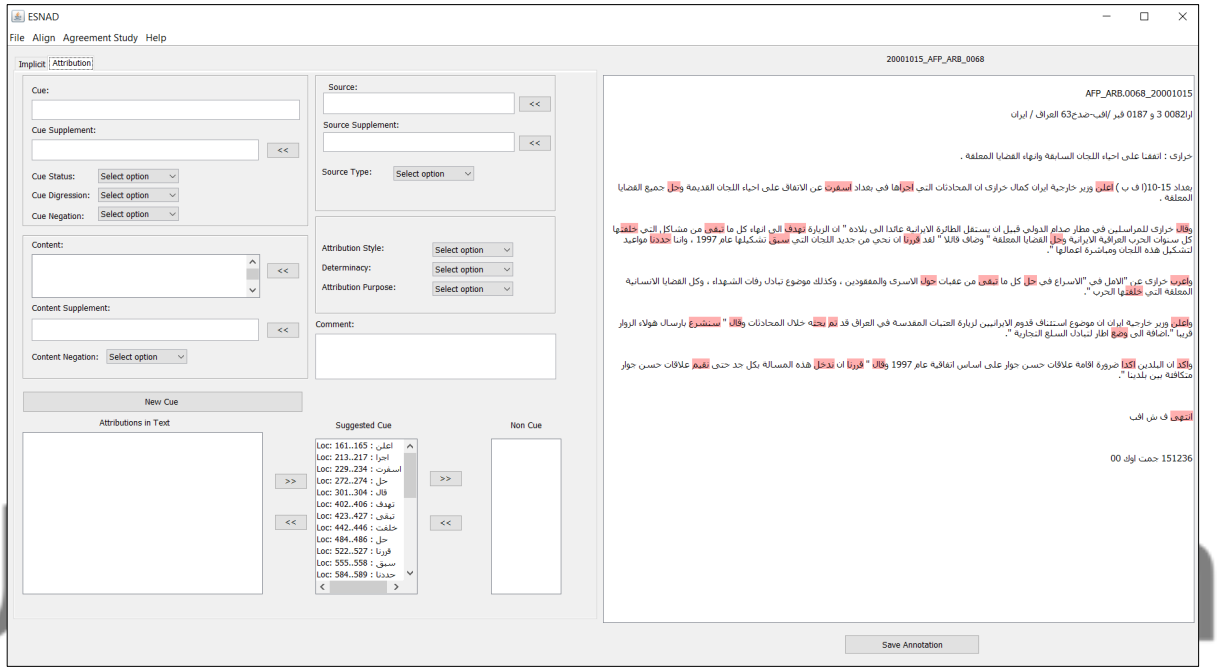

Figure 1: Main interface and initial state of ESNAD, where all potential cues are highlighted in the text and listed in the “Suggested Cue”

Inter-annotator Agreement

After annotating attribution relations in ATB Part 1, we applied the agreement study implemented by ESNAD. This function takes as arguments the annotated files from the first and second annotators. Five measures are applied: for label elements, we conduct observed agreement, precision, recall, F-score, and Cohen’s kappa (Siegel 1956) to consider any random or expected agreement by chance.

The level of observed agreement is calculated by dividing the number of instances agreed on by both annotators by the total number of instances considered. Precision and recall are two important measures, the weighted average of which is known as the F-score. Generally, precision refers to the proportion of correctly predicted positive instances relative to the overall predicted positive instances, while recall refers to the number of correctly predicted positive instances relative to the actual positive instances (Joshi 2016). Table 3 provides an overview of the results obtained by applying the inter-annotator agreement study of label features.

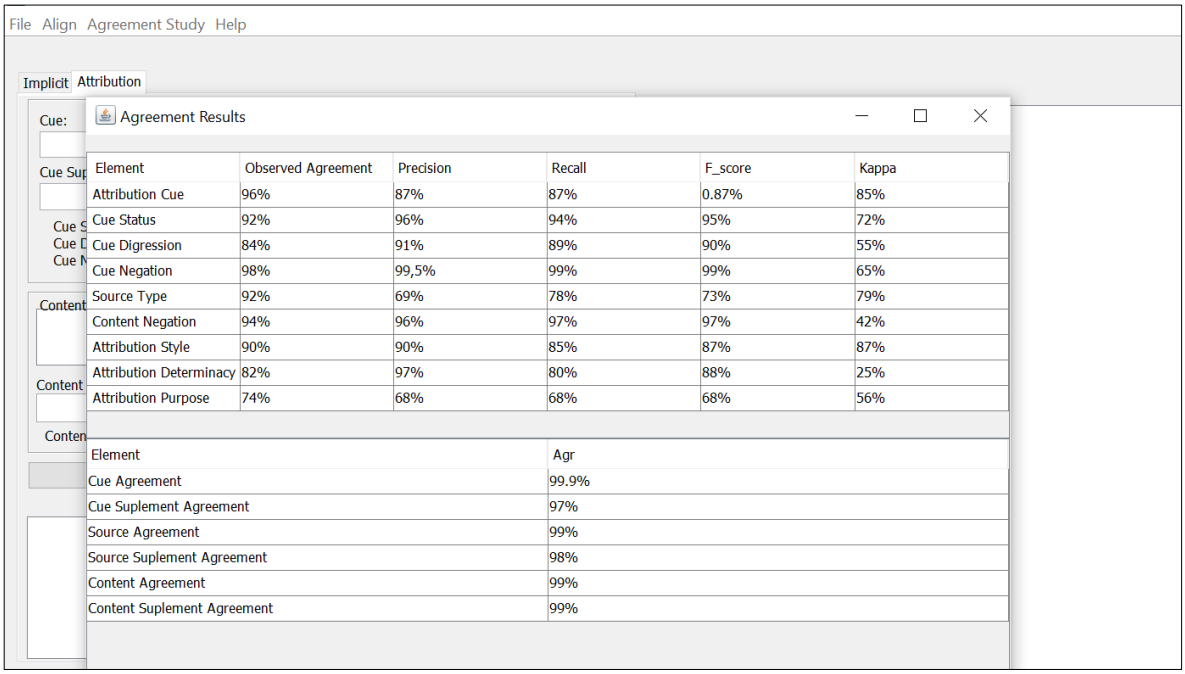

Figure 2: Screenshot of inter-annotator agreement study in ESNAD

Table 3: The results of inter-annotator agreement on all labeled features

Total potential attributions	13,260				
Total attribution relations	2,334				
Element	Observed Agreement	Precision	Recall	F-score	Kappa
IsCueAttribution	96%	87%	87%	87%	**85%**
Cue Status	92%	96%	94%	95%	**72%**
Cue Digression	84%	91%	89%	90%	55%
Cue Negation	98%	99.5%	99%	99%	65%
Source Type	92%	69%	78%	73%	**79%**
Content Negation	94%	96%	97%	97%	42%
Attribution Style	90%	90%	85%	87%	**79%**
Attribution Determinacy	82%	97%	80%	88%	25%
Attribution Purpose	74%	68%	68%	68%	56%

For text span elements such as attribution content, source, and cue supplement, we applied the Agr measure (Artstein and Poesio 2008). Exact Match is equal to 1 when exactly both annotators marked the same text span, whereas it is equal to 0 when there is no intersection between each annotator's selections. For partial agreement, we used the AGR measure presented in Equation (1), where Ann1 refers to the first annotator, Ann2 refers to the second annotator, and tokens refer to words. The overall agreement by

AGR is the average of both directions of AGR(Ann1||Ann2) and AGR(Ann2||Ann1). Table 4 presents all Agr measurements.

$$AGR(Ann1||Ann2) = \frac{(\#matched\ tokens\ between\ Ann1\ and\ Ann2)}{(\#tokens\ of\ Ann1)} \quad (1)$$

Table 4: Inter-annotator agreement of text features using *AGR* metric

Text Span Agreement (agreed attribution = 2,334)	*Avg Agr*
Cue Supplement Agreement	0.97
Source Agreement	0.99
Source Supplement Agreement	0.98
Content Agreement	0.99
Content Supplement Agreement	0.99

6. Results and Discussion

The annotation process was undertaken over nearly 6 months, including a training period and reviewing period afterward. As shown in Table 3, the annotators agreed on 2,334 attribution instances as attribution relations, with 189 distinct cues (Appendix A gives a subset of these cues). As described previously, we highlighted all verbs in the Arabic Treebank (ATB) and asked each annotator to decide on its function (i.e., whether it is a cue or not). The accuracy of *IsCueAttribution* was 96% with an F-score of 87% and a Cohen's kappa value of 85%.

The main reason for the disagreement in the results occurred when the cue is not a verb; noun or adverb cues were not highlighted in the tool. Therefore, the annotator may have overlooked implicit attributions of these cues, in particular. For instance, the cue (according/ AstnAdA/استنادا) is a noun cue of attribution that was missed by one of the annotators. Another case of disagreement was the confusion over considering the modifier function words as part of the cue or as part of the supplemental information of the cue, as in Example 17. Nested attribution was also sometimes missed by one of the annotators who only considered the higher level attribution. For example, the nested attribution located in the supplement of the source at the relative clause in Example 18 was missed by one annotator.

Ex.17
استنادا[لى] مصدر دبلوماسي فلسطيني {فى هانوي} *فان عرفات سيزور كلا من كوالا لمبور وجاكرتا وطوكيو.*

According [to] a Palestinian diplomatic source {in Hanoi}, *Arafat will visit Kuala Lumpur, Jakarta, and Tokyo.*
Cue Status: Explicit, Cue Purpose: Declaration, Attribution Style: Indirect, Determinacy: Yes
From: 20000815_AFP_ARB_0019.raw

Ex.18
و**اضاف** الشاب {الذي **قال** انه عنصر سابق من الجيش } *"ما هي الجرائم التي اقترفتها نساؤنا واطفالنا ليعيشوا حياة بائسة كلاجئين من بلد الى بلد.."*
The young man, {who **said** he was a former member of the Army}, **added**: *"What crimes did our women and children commit to lead such a miserable life as refugees from country to country?"*
Cue Status: Explicit, Cue Purpose: Declaration, Attribution Style: Direct, Determinacy: Yes
From: 20000815_AFP_ARB_0068.raw

Based on the results described above, we conducted the rest of the agreement measurements of the other elements on the 2,334 agreed attributions. The agreement on cue status (explicit or implicit) achieved 92% accuracy, 95% in F-score, and 72% in Cohen's kappa. Disagreements in this element arose from the ambiguity associated with determining the cue in implicit attribution. Certain cues are frequently used as explicit cues such as (announced/ >Eln/أعلن, said/qAl/قال , mentioned/*kr/ذكر), but it is not always explicit when the cue performs the function of the main verb in the content, as in Example 19.

Another ambiguity in the cue status was related to the existence of al-maSdar nominalization (i.e., where a noun functions semantically as a verb) after the attribution cue. Therefore, according to our guidelines, if there is only one main verb expressing the attribution and the

Ex.19
...**اعلن** [في تموز/يوليو *1998*] عن اجراء اول اختبار لهذا الصاروخ.

… ***announced*** *[in July 1998] that the first test of the missile would be carried out.*
Cue Status: Implicit, Cue Purpose: Declaration, Attribution Style: Indirect, Determinacy: Yes
From: 20000715_AFP_ARB_0072.raw

action/fact in content, then the cue status should be implicit. However, the existence of al-maSdar might confuse certain annotators, particularly regarding the issue of whether to consider only verbs as attributed cues. For example, al-maSdar occurs in Example 20, where the verb cue is (announced/أعلن) and the al-maSdar is (liberation/تحرير), which is a part of the content that operates as a verb. In this case, the annotator considered the cue as implicit despite the fact that al-maSdar introduces the content span.

Ex.20
اعلن مصدر رسمي *تحرير العناصر المحاصرين منذ اكثر من شهرين..*
An official source **announced** *the liberation of the trapped elements for more than two months …*
Cue Status: Implicit, Cue Purpose: Declaration, Attribution Style: Indirect, Determinacy: Yes
From: 20000715_AFP_ARB_0081.raw

Agreement on cue digression was 84% in terms of accuracy, 90% in terms of F-score, and 55% in terms of Cohen's kappa. The default status is "no", but the annotators may have been confused when the semantic function of the cue usually exhibits digression, such as (adding/mudifaan/مضيفا) and (pointing/mushiraan/مشيرا), which are frequently used to add content to the main attribution. Sometimes, they considered it as a new attribution with no digression, as in Example 21. Other disagreements resulted from the following: firstly, using different lexical terms to refer to the source in the additional attribution, as in Example 22, where (George/جورج) is (The new Al-Nasr coach /مدرب النصر الجديد) who is mentioned in the last attribution; and secondly, non-adjacent attributions of the same source, especially in news articles where a list of attributions are mentioned consecutively.

Ex.21

اضاف الشاهد *ان عسكريين طلبوا من سكان الحي الاحتفاظ بالهدوء والبقاء في منازلهم.*
The witness **added** *that the soldiers had asked the residents of the neighborhood to keep calm and stay in their homes.*
Cue Status: Explicit, Cue Digression: No, Cue Purpose: Assertion, Attribution Style: Indirect, Determinacy: Yes
From: 20000715_AFP_ARB_0070.raw

Ex.22
و**وعد** جورج [الجمهور السعودي] *بأنه "سيشاهد فريق النصر بثوب جديد مختلف عن المواسم الماضية من الناحية الفنية"* و**اضاف** مدرب النصر الجديد *"ان خبرتي في مجال التدريب تمتد الى عشرين عاما وابتعادي في الموسم الماضي كان لالتقاط الانفاس والبقاء مع عائلتي برغم انني تلقيت عروضا عدة من اندية فرنسية وايطالية واسبانية لكني فضلت الراحة".*
George **promised** [the Saudi public] *that he would "see the Al-Nasser team in a new outfit that is different from the previous seasons technically".* The new Al-Nasr coach **added**, *"My experience in the field of training extends to twenty years, and my separation last season was to catch my breath and stay with my family. Although I received several offers from French, Italian, and Spanish clubs, I preferred to rest".*
Cue Status: Explicit, Cue Digression: No, Cue Purpose: Promise, Attribution Style: Direct, Determinacy: Yes Cue Status: Explicit, Cue Digression: No, Cue Purpose: Assertion, Attribution Style: Direct, Determinacy: Yes
From: 20000715_AFP_ARB_0060.raw

High agreement was achieved on cue negation, with 98% accuracy, 99% F-score, and 65% Cohen's kappa. The low Cohen's kappa value can be attributed to the fact that most cues are not negated. The limited number of ambiguous cases arose due to confusion over the semantic function of the cue. In particular, one annotator focused on the semantic function of the cue itself, as in Example 23, while the other considered the meaning of the attribution, which conveys a degree of negation, see Example 24.

Ex.23
رفض مترجم السفارة المثول امام القضاء في اطار التحقيق في اعتداءات وقعت في 1986 في باريس.
The embassy interpreter ***refused*** *to appear in court as part of an investigation into the 1986 attacks in Paris.*
Cue Status: Implicit, Cue Negation: Yes, Cue Purpose: Assertion, Attribution Style: Indirect, Determinacy: Yes
From: 20000715_AFP_ARB_0080.raw

Ex.24
أعلن سعيد انه سيرحل اذا لم يحقق الاهلي مطالبه التي تؤمّن مستقبله.
Saeed **announced** *that he would leave if Al-Ahly did not fulfill his demands that would secure his future***.**
Cue Status: Explicit, Cue Negation: No, Cue Purpose: Declaration, Attribution Style: Indirect, Determinacy: No
From: 20000815_AFP_ARB_0022.raw

Regarding the cue supplement, which is any additional information related to the cue itself, the level of agreement amounted to 97% on the AGR metric. One annotator included the modifier in the cue and the other included it in the cue supplement, as in Example 17. Some supplemental information occurs far from the cue itself, which led one annotator to miss marking it as a cue supplement or to include it in other supplemental features (e.g., for content or source).

For source, a high level of agreement was achieved (99%) on the source span, which presented a reliable description in our guidelines. However, the source type (WR, EXP-AG, or IMP-AG) achieved an accuracy of 92%, an F-score of 73%, and a Cohen's kappa of 79%. Although the feature was defined clearly in the scheme, disagreements occurred between EXP-AG and IMP-AG. In particular, the position of the source in the attribution was not close enough to the cue or it was mentioned in the previous

attribution with digression attributions. In Example 25, the source of the second attribution was mentioned in the first attribution. One annotator considered the source here as IMP-AG but the other marked it as EXP-AG since the source is in the same sentence span. For the source supplement, which consists predominantly of relative clauses describing the source, a high level of agreement (98%) was achieved using the AGR metric. The disagreements that did occur may have been caused by the inclusion of the description before the source in the supplement or the source itself.

Ex.25
اعترف *مدرب يوفنتوس الدولي السابق كارلو انشيلوتي بصعوبة المباراة، **وقال** "لقد وجدنا صعوبة في فرض اسلوب لعبنا بسبب حالة ارضية الملعب اولا وقوة باري الذي خلق لنا متاعب عدة".*
*Former Juventus international coach Carlo Ancelotti **admitted** the difficulty of the match*, and **said**, *"We found it difficult to impose our style of play because of the condition of the pitch first and Barry's strength, which created many troubles for us."*
Cue Status of (admitted/اعترف): Implicit, Cue Purpose: Declaration, Attribution Style: Indirect, Determinacy: Yes Cue Status of (said/قال): Explicit, Cue Purpose: Declaration, Attribution Style: Direct, Determinacy: Yes
From: 20001015_AFP_ARB_0135.raw

In content boundary annotation, 99% agreement was achieved on the AGR metric. Unsurprisingly, this high agreement is due to the fact that most attribution in our corpus is explicit, with exact words marked with quotations, or implicit, where the content should contain the full sentence (including the cue). For content negation, most disagreements were caused by the inclusion of the cue in the content in an implicit attribution (as proposed in the scheme), especially when the cue was negated, such as (deny/ nfY /نفى) and (refuse/ rfD/رفض). In turn, this led one annotator to consider it as negated content. In addition, it is sometimes the case that an indirect attribution style creates ambiguity over the content boundaries and the content negation, as in Example 26. The content supplement is also clear enough in our scheme, where the agreement is 99%. Limited supplemental information was missed because the annotators included it in the content for indirect attributions. Example 27 shows one case of disagreement, where the sentence (which has a range of 1,300 kilometers/ الذي يصل مداه الى 1300 كيلومتر) was marked as content instead of content supplement.

Measurements were also taken of the level of agreement on the general features of attribution purpose, attribution determinacy, and attribution style. The annotators agreed on the attribution purpose at the levels of 74%, 68%, and 56% in terms of accuracy, F-score, and Cohen's kappa,

Ex.26
اضاف *انه ليس متاكدا من ان يشارك على الفور احد كبار المسؤولين الاميركيين في هذه اللقاءات .*
He **added** *that he was not sure whether a senior American official would participate in these meetings immediately.*
Cue Status: Explicit, Cue Negation: No, Cue Purpose: Declaration, Attribution Style: Indirect, Content Negation: Yes, Determinacy: Yes
From: 20000915_AFP_ARB_0073.raw

Ex.27
اعلن الجيش *انه نفذ "بنجاح" اليوم السبت اختبارا جديدا للصاروخ (الذي يصل مداه الى 1300 كيلومتر) .*
The army **announced** *that it had "successfully" carried out today, Saturday, a new test of the missile* (which has a range of 1,300 kilometers).
Cue Status: Explicit, Cue Purpose: Declaration, Attribution Style: Indirect, Determinacy: Yes
From: 20000715_AFP_ARB_0040.raw

respectively. Although we relied on the well-known Speech Act Theory when labeling the purpose of the attribution (using the semantic function of the cue), we found that a large group of speech acts were ambiguous in this empirical annotation. For example, the cues (said/qal/قال) and (added/>DAf/أضاف), which are often used for Declaration purposes, are also used commonly for Assertion purposes in a specific context. Appendix A shows the most common cues and their usage percentage in our corpus. Many instances of disagreement were also identified for attribution purpose on either Assertion or Expression, especially for cues with both functions. For instance, in Example 28, some information in the content supplement plays a role in distinguishing the purpose of the attribution itself, which is also the

case in Example 29. We also noticed the ambiguity associated with attribution purpose in the pilot annotation, and so we defined a decision tree to assist the annotator in examining attribution purposes from the less ambiguous to the more ambiguous. This minimized the probability of selecting the most ambiguous purposes in the first place. Our findings here will re-establish linguistic studies of the purposes of speech acts using real-world examples.

Ex.28
وزيرة النقل [على الفور] العملية ووصفتها بانها "عمل جبان قامت به عناصر معادية تملكها اليأس". **أدانت**
The Minister of Transport ***condemned*** *[immediately] the operation and she described it as "a cowardly act committed by hostile elements full of despair."*
Cue Status: Implicit, Cue Purpose: Assertion, Attribution Style: Indirect, Determinacy: Yes
From: 20000815_AFP_ARB_0002.raw

Ex.29
"لقد وجدنا صعوبة في فرض اسلوب لعبنا بسبب حالة ارضية الملعب اولا وقوة باري الذي خلق لنا متاعب عدة". **قال**
He **said,** *"We found it difficult to impose our style of play because of the condition of the pitch first and Barry's strength, which created many troubles for us."*
Cue Status: Explicit, Cue Purpose: Expression, Attribution Style: Direct, Determinacy: Yes
From: 20001015_AFP_ARB_0213.raw

In attribution determinacy, the level of agreement was only 82%. It is worth noting that the negation features (cue negation and content negation) may have undermined the annotators' correct decisions about attribution determinacy. This feature may require a strong linguistic background, which was not applicable to our annotators. In contrast, the level of agreement was high in terms of attribution style, which achieved 90% accuracy, 87% in F-score, and 79% in Cohen's kappa. In this case, disagreements mainly occurred with indirect attribution when using the particle (that/ >n/أن) with no quotation marks. In such instances, the annotator may regard the attribution style as direct since the content is almost exactly the same in terms of the lexical items used.

7. The AttImam Corpus Distribution

By examining annotation distribution in a corpus in Table 5, we identified a large variety of distinct cues (189), most of which were explicit cues (77%). The levels of direct and indirect usage were 36% and 64%, respectively. These distributions fit with the domain of the corpus (i.e., news articles), in which the writer tends to use words in reporting speech with high determinacy (94.17%) and non-negation (2.44%) in terms of cue negation. The results indicate that 25% of the attributions are related to previous attributions, which is relevant to cue digression. Moreover, the news articles contained multiple types of supplemental information, and they explicitly mentioned the source (with 75% of EXP-AG in source type). This is consistent with the desire of journalists to increase the truthiness of the news.

The most common attribution purposes were Declaration (1,345, 57.6%), Assertion (647, 27.7%), and Directive (164, 7.2%). In addition, the most commonly used cues for Declaration were (added/اضاف, announced/أعلن, and transferred/نقل); for Assertion, the most commonly used cues were (added/اضاف, confirmed/أكد, said/قال); and finally, (call/دعا, asked/طالب) was the most commonly used cue in the Directive attribution purpose.

Table 5: Annotation distribution in Arabic attribution corpus

Attribution Instances	2,334
Distinct Cues	189
Explicit Cues	1799 (77%)
Implicit Cues	535 (23%)
Common Purposes	Declaration (1,345, 57.6%), Assertion (647, 27.7%), Directive (164, 7.2%), Expression (156, 6.68%), Promise (17), Questioning (5)
Commonly Used Cues	say/qAl/قال(376), declare/A'gn/أعلن(294), add/ATaf/ ,(232)أضاف ذكر,(133)أكد,أوضح (155)(114)
Supplements	Cue (920), source (335), content (95)
Source Type	EXP_AG (1746, 75%), IMP_AG (489, 21%), Miss (80,3.43%), and WR (19,0.18%)
Cue Negation	57 cues with negation (2.44%)
Cue Digression	588 cues with digression (25%)

Content Negation	167 instances with content negation (7.16%)
Attribution Style	Direct at 842 (36%), Indirect at 1,492 (64%)
Attribution Determinacy	2,198 determined attributions (94%)

8. Conclusion

The ImamAtt TB corpus is the first empirical effort to construct a gold standard corpus of attribution in Arabic. With this corpus, researchers in diverse fields such as authorship identification, news validation, and sentiment and opinion analysis can extract lexical, morphological, syntactic, semantic, discourse, and attribution features. We report new linguistic findings regarding writing styles and attribution purposes, which will benefit linguistic research in Arabic. This corpus will be used to develop machine learning models to identify attribution elements and boundaries automatically for new pieces of text. The corpus will be distributed via LDC soon.

Acknowledgments

I am grateful to the Deanship of Scientific Research at Imam University for funding this research. I would also like to express my thanks to all of the consultants – in both linguistics and language modeling – who contributed their expertise to this paper, including Dr. Fatmah Alshehriy, Prof. Bonee Webber, and Dr. Jehan. A very special thanks to the annotators, too, who were patient throughout the long discussion and annotation process: Nora Alfayz, Sarah Alkaran, Rawan Alharbi, Ruba Alhamed, Munera Alshamari, Duha Alqahtani and Hessah Alkalaf.

References

Ahmed, Al-Falahi, Ramdani Mohamed, and Bellafkih Mostafa. 2019. "Arabic Poetry Authorship Attribution Using Machine Learning Techniques." *Journal of Computer Science* 15(7): 1012–21.

Al-Ayyoub, Mahmoud, Ahmed Alwajeeh, and Ismail Hmeidi. 2017. "An Extensive Study of Authorship Authentication of Arabic Articles." *International Journal of Web Information Systems* 13(1).

Al-Saif, Amal et al. 2018. "Annotating Attribution Relations in Arabic." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*,.

Al-Saif, Amal, and Katja Markert. 2010. "The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic." In *LREC*.

Al-Sarem, Mohammed et al. 2020. "Ensemble Methods for Instance-Based Arabic Language Authorship Attribution." *IEEE Access* 8: 17331–45.

Alam, Hassan, and Aman Kumar. 2013. "Multi-Lingual Author Identification and Linguistic Feature Extraction—A Machine Learning Approach." In *2013 IEEE International Conference on Technologies for Homeland Security (HST)*, 386–89.

Almeida, Mariana S C, Miguel B Almeida, and André F T Martins. 2014. "A Joint Model for Quotation Attribution and Coreference Resolution." In *EACL*, 39–48.

Altheneyan, Alaa Saleh, and Mohamed El Bachir Menai. 2014. "Naïve Bayes Classifiers for Authorship Attribution of Arabic Texts." *Journal of King Saud University-Computer and Information Sciences* 26(4): 473–84.

Alwajeeh, Ahmed, Mahmoud Al-Ayyoub, and Ismail Hmeidi. 2014. "On Authorship Authentication of Arabic Articles." In *2014 5th International Conference on Information and Communication Systems (ICICS)*, 1–6.

Artstein, Ron, and Massimo Poesio. 2008. "Survey Article Inter-Coder Agreement for Computational Linguistics." (December 2007).

Austin, John L. 1975. "How to Do Things with Words (JO Urmson & M. Sbisa, Eds.)." *Harvard U. Press, Cambridge, MA*.

Badawi, El Said, Michael Carter, and Adrian Gully. 2013. *Modern Written Arabic: A Comprehensive Grammar*. Routledge.

El Bakly, Abeer H, Nagy Ramadan Darwish, and Hesham A Hefny. 2020. "Using Ontology for Revealing Authorship Attribution of Arabic Text." *International Journal of Engineering and Advanced Technology (IJEAT)* 9(4): 143–51.

BwEyAd, NwArp, and Emr Blxyr. 2012. "TSnyf OfEAl AlklAm Fy AlxTAb AlSHAfy AljzAry Almktwb BAllgp AlErbyp." *mjlp AlOvr AlOdbyp*.

Carlson, Lynn, and Daniel Marcu. 2001. "Discourse Tagging Reference Manual." *ISI Technical Report ISI-TR-545* 54: 56.

Cheng, Na, Rajarathnam Chandramouli, and K P Subbalakshmi. 2011. "Author Gender Identification from Text." *Digital Investigation* 8(1): 78–88.

Dinesh, Nikhil et al. 2005. *Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives*. http://www.cis.upenn.edu/pdtb. (January 9, 2021).

Elson, David K, and Kathleen McKeown. 2010. "Automatic Attribution of Quoted Speech in Literary Narrative." In *AAAI*, Citeseer.

Fernandes, William Paulo Ducca, Eduardo Motta, and Ruy Luiz Milidiú. 2011. "Quotation Extraction for Portuguese." In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL 2011), Cuiabá*, 204–8.

Guzmán-Cabrera, Rafael, Manuel Montes-y-Gómez, Paolo Rosso, and Luis Villasenor-Pineda. 2009. "Using the Web as Corpus for Self-Training Text Categorization." *Information Retrieval* 12(3): 400–415.

Habash, Nizar Y. 2010. "Introduction to Arabic Natural Language Processing." *Synthesis Lectures on Human Language Technologies* 3(1): 1–187.

Ide, Nancy, and Keith Suderman. 2009. "Bridging the Gaps: Interoperability for GrAF, GATE, and UIMA." In *Proceedings of the Third Linguistic Annotation Workshop*, Association for Computational Linguistics, 27–34.

Joshi, Renuka. 2016. "Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures - Exsilio Blog." https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/ (September 9, 2016).

Juola, Patrick, and R Harald Baayen. 2005. "A Controlled-Corpus Experiment in Authorship Identification by Cross-Entropy." *Literary and Linguistic Computing* 20(Suppl): 59–67.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 2009. "Computational Methods in Authorship Attribution." *Journal of the American Society for information Science and Technology* 60(1): 9–26.

Mamede, Nuno, and Pedro Chaleira. 2004. "Character Identification in Children Stories." In *Advances in Natural Language Processing*, Springer, 82–90.

Mann, William C, and Sandra A Thompson. 1988. "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization." *Text-Interdisciplinary Journal for the Study of Discourse* 8(3): 243–81.

Matloob, A. 1986. "A Dictionary of Rhetorical Terms and Their Development, Part 2, Taa'-Khaa'." *Baghdad: The Iraqi Scientific Academy*.

Müller, Christoph, and Michael Strube. 2006. "Multi-Level Annotation of Linguistic Data with MMAX2." *Corpus technology and language pedagogy: New resources, new tools, new methods* 3: 197–214.

Neumann, Hendrik, and Martin Schnurrenberger. 2009. "E-Mail Authorship Attribution Applied to the Extended Enron Authorship Corpus (XEAC)."

O'Keefe, Tim et al. 2012. "A Sequence Labelling Approach to Quote Attribution." In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, 790–99.

Ombabi, Abubakr H, Wael Ouarda, and Adel M Alimi. 2020. "Deep Learning CNN--LSTM Framework for Arabic Sentiment Analysis Using Textual Information Shared in Social Networks." *Social Network Analysis and Mining* 10(1): 1–13.

Otoom, Ahmed Fawzi et al. 2014. "An Intelligent System for Author Attribution Based on a Hybrid Feature Set." *International Journal of Advanced Intelligence Paradigms* 6(4): 328–45.

Ouamour, Siham, and Halim Sayoud. 2018. "A Comparative Survey of Authorship Attribution on Short Arabic Texts." In *International Conference on Speech and Computer*, 479–89.

Pareti, Silvia. 2011. "Annotating Attribution Relations and Their Features." In *Proceedings of the Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval*, ACM, 19–20.

Pareti, Silvia. 2012. "A Database of Attribution Relations." In *LREC*, , 3213–17.

Pareti, Silvia et al. 2013. "Automatically Detecting and Attributing Indirect Quotations." In *EMNLP*,

989–99.

Pareti, Silvia. 2016. “PARC 3.0: A Corpus of Attribution Relations.” In *LREC*, Portoroz.

Pareti, Silvia, and Irina Prodanof. 2010. “Annotating Attribution Relations: Towards an Italian Discourse Treebank.” In *LREC*.

Pillay, Sangita R, and Thamar Solorio. 2010. “Authorship Attribution of Web Forum Posts.” In *ECrime Researchers Summit (ECrime), 2010*, IEEE, 1–7.

Pouliquen, Bruno, Ralf Steinberger, and Clive Best. 2007. “Automatic Detection of Quotations in Multilingual News.” In *Proceedings of Recent Advances in Natural Language Processing*, 487–92.

Prasad, Rashmi et al. 2006. “Attribution and Its Annotation in the Penn Discourse TreeBank.” *TAL* 47(2): 43–64.

Prasad, Rashmi et al. 2007. “The Penn Discourse Treebank 2.0 Annotation Manual.”

Pustejovsky, James, and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. “ O’Reilly Media, Inc.”

Rabab’ah, Abdullateef, Mahmoud Al-Ayyoub, Yaser Jararweh, and Monther Aldwairi. 2016. “Authorship Attribution of Arabic Tweets.” *AICCSA* .

Sarmento, Luis, Sergio Nunes, and E Oliveira. 2009. “Automatic Extraction of Quotes and Topics from News Feeds.” In *4th Doctoral Symposium on Informatics Engineering*,.

Searle, John R. 1976. “A Classification of Illocutionary Acts.” *Language in society* 5(1): 1–23.

Siegel, Sidney. 1956. Nonparametric statistics for the behavioral sciences. *Nonparametric Statistics for the Behavioral Sciences.* New York, NY, US: McGraw-Hill.

Van Son, Chantal et al. 2016. “GRaSP: A Multilayered Annotation Scheme for Perspectives.” In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 1177–84.

Stenetorp, Pontus et al. 2012. “BRAT: A Web-Based Tool for NLP-Assisted Text Annotation.” In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 102–7.

Tan, Richmond Hong Rui, and Flora S Tsai. 2010. “Authorship Identification for Online Text.” In *Cyberworlds (CW), 2010 International Conference On*, IEEE, 155–62.

Wiebe, Janyce. 2002. “Instructions for Annotating Opinions in Newspaper Articles.”

Wolf, Florian, and Edward Gibson. 2005. “Representing Discourse Coherence: A Corpus-Based Study.” *Computational Linguistics* 31(2): 249–87.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. “WebAnno: A Flexible, Web-Based and Visually Supported System for Distributed Annotations.” In *ACL (Conference System Demonstrations)*, 1–6.

Ywns, Ely mHmd mHmd. 2004. “Mqdmp Fy Elmy AldlAlp w AltxATb.”

Appendix A. The cue distribution

The following table shows the frequency of the most common cues for each syntactic type of attributed span. For the complete list contact the authors.

Verbal Cues		
Cue	**Frequency**	**Additional features**
said/qal/قال	285	Cue Status (Explicit (284), Implicit (1)); Attribution Purpose (Assertion (71), Declaration (205), Directive (2), Expression (5))
declare/ >Eln/أعلن	204	Cue Status (Explicit (164), Implicit (40)); Attribution Purpose (Assertion (16), Declaration (186), Directive (1), Expression (1))
added/ ADAf/اضاف	181	Cue Status (Explicit (179), Implicit (2)); Attribution Purpose (Assertion (49), Declaration (123), Directive (2), Expression (7))
explained/ AwDH/اوضح	103	Cue Status (Explicit (102), Implicit (1)); Attribution Purpose (Assertion (21), Declaration (81), Directive (1))
confirmed/ Akd/اكد	100	Cue Status (Explicit (86), Implicit 4); Attribution Purpose (Assertion (97), Declaration (3))
Noun Cues		
Cue	**Frequency**	**Additional features**
سؤالquestion/ s&Al/	3	Cue Status (Explicit (2), Implicit (1)); Attribution Purpose (Questioning (3))
استناداbased on/ AstnAdA/	3	Cue Status (Explicit (3)); Attribution Purpose (Declaration (3))
طلبهrequest/ Tlbh/	2	Cue Status (Implicit (2)); Attribution Purpose (Directive (2))
قولهsay/ qwlh/	2	Cue Status (Explicit (2)); Attribution Purpose (Declaration (2))
اشارةreferring/ A$Arp/	1	Cue Status (Implicit (1)); Attribution Purpose (Assertion (1))
Cues of prepositional phrase		
Cue	**Frequency**	**Additional features**
according to/bHsb /بحسب	8	Cue Status (Explicit (5), Implicit (3)); Attribution Purpose (Assertion (2), Declaration (6))
Scheduled/mn Almqrr/من المقرر	3	Cue Status (Explicit (3)); Attribution Purpose (Assertion (2), Declaration (1))
according to/ wbHsb /وبحسب	2	Cue Status (Implicit (2)); Attribution Purpose (Assertion (1), Declaration (1))
بقولهby saying/ bqwlh/	1	Cue Status (Explicit (1)); Attribution Purpose (Declaration (1))
by reference/ bAlAstnAd/بالاستناد	1	Cue Status (Implicit (1)); Attribution Purpose (Declaration (1))

Implicit Cue	
Frequency	**Additional features**
7	Cue Status (Explicit (6), Implicit (1)); Attribution Purpose (Assertion (1), Declaration (6))