Attribution: A Computational Approach

Silvia Pareti

Doctor of Philosophy

Institute for Language, Cognition and Computation

School of Informatics

University of Edinburgh

2015

Abstract

Our society is overwhelmed with an ever growing amount of information. Effective management of this information requires novel ways to filter and select the most relevant pieces of information. Some of this information can be associated with the source or sources expressing it. Sources and their relation to what they express affect information and whether we perceive it as relevant, biased or truthful. In news texts in particular, it is common practice to report third-party statements and opinions. Recognizing relations of attribution is therefore a necessary step toward detecting statements and opinions of specific sources and selecting and evaluating information on the basis of its source.

The automatic identification of Attribution Relations has applications in numerous research areas. Quotation and opinion extraction, discourse and factuality have all partly addressed the annotation and identification of Attribution Relations. However, disjoint efforts have provided a partial and partly inaccurate picture of attribution. Moreover, these research efforts have generated small or incomplete resources, thus limiting the applicability of machine learning approaches. Existing approaches to extract Attribution Relations have focused on rule-based models, which are limited both in coverage and precision.

This thesis presents a computational approach to attribution that recasts attribution extraction as the identification of the attributed text, its source and the lexical cue linking them in a relation. Drawing on preliminary data-driven investigation, I present a comprehensive lexicalised approach to attribution and further refine and test a previously defined annotation scheme. The scheme has been used to create a corpus annotated with Attribution Relations, with the goal of contributing a large and complete resource than can lay the foundations for future attribution studies.

Based on this resource, I developed a system for the automatic extraction of attribution relations that surpasses traditional syntactic pattern-based approaches. The system is a pipeline of classification and sequence labelling models that identify and link each of the components of an attribution relation. The results show concrete opportunities for attribution-based applications.

Lay Summary

Our society is overwhelmed with an ever growing amount of information. Effective management of this information requires novel ways to filter and select the most relevant pieces of information. Some of this information can be associated with the source or sources expressing it. Sources and their relation to what they express affect information and whether we perceive it as relevant, biased and/or truthful. In news texts in particular, it is common practice to report third-party statements and opinions. Recognizing relations of attribution is therefore a necessary step toward detecting statements and opinions of specific sources and selecting and evaluating information on the basis of its source.

The automatic identification of Attribution Relations has applications in numerous research areas. Quotation and opinion extraction, discourse and factuality have all partly addressed the annotation and identification of Attribution Relations. However, disjoint efforts have provided a partial and partly inaccurate picture of attribution. Moreover, these research efforts have generated small or incomplete resources, thus limiting the applicability of machine learning approaches. Existing approaches to extract Attribution Relations have focused on rule-based models, which are limited both in coverage and precision.

This thesis presents a computational approach to attribution that recasts attribution extraction as the identification of the attributed text, its source and the lexical cue linking them in a relation. Drawing on preliminary analysis of attribution, I present a comprehensive approach to attribution and further refine and test a previously defined annotation scheme. The scheme has been used to create a corpus annotated with Attribution Relations, with the goal of contributing a large and complete resource that can lay the foundations for future attribution studies.

Based on this resource, I developed a system for the automatic extraction of attribution relations that surpasses traditional hand-crafted approaches. The system comprises different components that learn directly from the data and are subsequently applied to identify each of the constitutive elements of an attribution relation. The results show concrete opportunities for applications such as the identification of quotations and opinions and the selection of information based on its source.

Acknowledgements

I would like to start by thanking my supervisor Bonnie Webber for giving me the chance to come to Edinburgh as an intern and for taking me back as a PhD candidate. Her guidance and support throughout these years was invaluable. I am sincerely thankful to my review committee, Rashmi Prasad, Beatrice Alex and Jon Oberlander, for their precious suggestions and reflections. I would also like to thank my second supervisor Kees van Deemter for his helpful advice and for guiding me towards semantic reflections. I am also grateful to Yannis Konstas for being like another supervisor to me. I knew I could always count on his sharp suggestions and feedback, in particular on technical issues.

I also wish to thank Mirella Lapata and Beatrice Alex and Jon Oberlander for their useful advice on my first year report and their helpful feedback after my DDD. Many thanks to all my colleagues in the Informatics Forum for being a great and inspiring company during my PhD studies. In particular, I would like to thank Diego Frassinelli, Eva Hasler, Desmond Elliot, Stella Frank and my brother Paolo Pareti. They truly made the PhD journey a memory to treasure.

This reaserch would not have been possible without the financial assistance of SICSA and the Erasmus Placement programme, which enabled me to recruit linguist annotators. I would like to thank my annotators Lorenzo Bernardini, Alessandra Cervone and Alice Bracchi for their time.

I am grateful to my family, for teaching me to learn and for spending so many hours on the phone with me when my tired voice was all I had to offer. My deepest thanks go to my husband Gregor Sarrazin, the loving man who let me go to follow my path but never left my side in spite of the distance, pushing and pulling me through every difficult time and sharing with me all the joys.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Silvia Pareti*)

To Dario.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

With a vast amount of data being available, in particular through the world wide web, more than ever before users have the chance to access an enormous amount of information. While information per se is a resource, this information overload can hinder our ability to process it and use it to understand issues or make decisions. To manage this vast amount of information requires ways to organise, filter and select it. It therefore becomes important to recognise different point of views (e.g. to make a medical or financial decision), monitor the statements of a specific person (e.g. a politician) and identify truthful and reliable information. These tasks require the identification of attribution relations, thus allowing to link the attributed material to the entity representing its source.

An immediate benefit of attributions is the possibility to identify what has been attributed to a specific source. Moreover, attributions affect how the text itself is perceived. Different sources and attribution choices have an impact on the interpretation and perception of the attributed material. Changes in the source or attributional verb can affect our perception of the quoted statement as illustrated in Ex. (1a), Ex. (1b) and Ex. (1c).

(1) a. Dr. Smith said: “There is no correlation between smoking cigarettes and the incidence of lung cancer in the population.”

b. Dr. Smith jokes: “There is no correlation between smoking cigarettes and the incidence of lung cancer in the population.”

c. A smoker said: “There is no correlation between smoking cigarettes and the incidence of lung cancer in the population.”

While research and commercial systems for the automatic identification and extraction of attribution relations have multiplied in recent years, several issues are still to be addressed. The applications of such systems are severely limited by low precision and low recall.

The reason for this relatively poor performance is partly to be found in the limited scope of such approaches. Studies on attribution focused either on its overlap and interaction with other linguistic phenomena, such as discourse relation and factuality, or on specific types of attributions, such as inter-sentential, direct quotations or having a Named Entity (NE) source. While these studies show that attribution is relevant for different linguist fields, their approaches address only a subset of attribution and rely on small and partially annotated resources. These resources are inadequate to guide a comprehensive understanding of attribution and drive the development of extraction systems.

Lacking a large annotated resource, the literature has so far produced only small scale studies, driven by assumptions based on intuition rather than statistically motivated. Since the lack of annotated data hindered the development of supervised computational models, the systems developed had to rely mostly on hand-crafted rules and results could be tested on a small number of examples.

This thesis proposes a computational approach to attribution that takes into account different types of attribution and the many ways it can be expressed. Drawing from a mosaic of theoretical and practical approaches, this work proposes to answers the following questions:

1. Is attribution a class of relations that, although not homogeneous, share fundamental characteristics and can be addressed as a whole, independently from other linguistic phenomena?

2. If this is the case, is it possible to consistently annotate a resource with a wide range of attribution relations?

3. If a large annotated resource is available, can attribution relations be automatically extracted with higher precision and recall than current rule-based systems by using supervised machine learning algorithms?

1.2 What are Attribution Relations

In previous work (Pareti, 2009), I proposed to deconstruct Attribution Relations (ARs) into three main elements:[1]

- *Content*: the linguistic material that is attributed, usually to a third party.
- **Source**: the entity the material is attributed to.
- Cue: the link that connects source and content in a relation of ownership, expressing a certain attitude.

I also suggested treating attribution as a textual relation whose constitutive elements can be identified as the text spans expressing them. While we can identify the content with a portion of text, it can be argued whether considering source and cue as lexically expressed is also a valid approach.

A different approach comes from the Speaker Attribution literature (see Sec. 2.2.1) which deals with a subset of attribution, namely the attribution of quotations. Quotation sources are identified as the entities that uttered the quoted material. Thus, coreference and anaphora resolution, but also entity resolution, are applied to mentions in order to identify the entity a mention refers to.

While we ultimately want to also identify the entity a mention refers to, in particular in case the source is pronominally expressed, this is not always useful or sufficient. Examples of entities which are less informative than their mention are illustrated in Ex. (2). In Ex. (2a) it would be less informative to retrieve the name of the spokewoman, since what is relevant is her relation with ‘Lorillard’. The same entity could be in another context the ‘witness’ of Ex. (2b), where the stress is on the relation between the entity and the event. In Ex. (2c) it would be very detrimental to just consider the mention ‘those’ and try to resolve it to an entity, as this would miss the partitive which is a key element to a correct comprehension of the AR.

(2) a. **A Lorillard spokewoman** said [...]

b. **A witness** described [...]

c. **50% of those interviewed** replied [...]

[1] Source, cue and content are identified in the examples in this thesis as follows: the text span corresponding to the source mention is bold, the font for the text corresponding to the cue element is underlined and the content is in italics.

In order to preserve the informativeness of the source, I propose to identify the source of an AR as the text span expressing it, which includes the entity mention and relevant modifiers. Coreference, anaphora and entity resolution are additional steps, complementary, but independent from attribution.

The cue element should also be lexically anchored. If no lexical element (e.g. an opinion verb or reporting punctuation) explicitly signals the AR, an AR cannot be established. While we can infer that something is not directly the opinion or the words of the author of a text, without an explicit AR cue, the author is presenting those words as her own.

1.2.1 A Complex Relation

What makes attribution such a complex relation and an unsolved challenge for attribution extraction tasks? This section will present an overview of the characteristics of ARs by drawing from previous qualitative analysis of the ways this relation is expressed in English and Italian (Pareti, 2009).

The complexity of attribution is partly due to the rich variety of expressions encoding it that makes the definition of a predictive structure not viable. The *content* can be expressed by as little as a single word, as in Ex. (3a). This includes the cases when the content is expressed by a pronoun or event anaphora as in Ex. (3b), where *it* refers to the previous unattributed quote.

(3) a. *"Sì"*, le risponde convinta **unamichetta**. (ISST cs060)[2]

"Yes", answers to her confident **a friend**.

b. "[. . .]". A dir*lo* è **Giuseppe Signori**, . . . (ISST re126)

"[. . .]". It is **Giuseppe Signori** to say *it*, . . .

More often the content is expressed by a clause or a group of clauses. Commonly this is the direct object of an attributional verb and the attributing span is the main clause, as in Ex. (4), or the content itself constitutes the main clause and the attributing span is parenthetically expressed, as in Ex. (5).

(4) **The assistant HHS secretary** said *the ban "should be continued indefinitely."* (wsj_0174)[3]

[2] Example from the Italian Syntactic Semantic Treebank corpus of newspaper articles (ISST) (Montemagni et al., 2003).

[3] Example from the Wall Street Journal corpus. The notation reflects the original file name.

(5) *Other airlines would have access to the system*, **they** said, *and negotiations with partners were already under way*. (wsj_1850)

The content of an attribution can also span over several consecutive sentences such as in Ex. (6). In some cases, particularly in interviews or testimonies, the content can be expressed discontinuously as a chain of attributions to the same source where only the first content is explicitly attributed.

(6) But *"the concept is workable. You sell the good bank as an ongoing operation and use some of the proceeds to capitalize the bad bank,"* says **thrift specialist Lewis Ranieri of Ranieri Associates in New York**. (wsj_0179)

There is also a high degree of freedom concerning the elements expressing the role of **source**, the other key component of attribution. Commonly, the source has the thematic role of experiencer of a private state or of agent of a speech event, and it is considered to be a NE. However, sources cannot always be identified through NE recognition and by applying anaphora resolution to resolve nominal and pronominal mentions of a NE.

In addition to specific named individuals such as 'Charles Bradford' or institutions such as 'Stewart & Stevenson Services Inc.', sources can also be not-named entities such as 'scientists', 'a witness' or 'the people' and both animate or inanimate, namely metonymic referents of the animate source producing them (e.g. 'newspaper', 'report', 'speech').

In addition, sources can be left implicit, namely in passive constructions or in pro-drop languages like Italian, or be completely omitted and not appear in the text. The latter attribution can still serve the purpose of removing liability from the writer when presenting information of uncertain origin or it can convey that it should be perceived as shared knowledge. Concealing the source can be achieved by means of verb structures not requiring a subject such as infinitival or passive forms as in Ex. (7), or by means of cues other than verbs as in Ex. (10).

(7) It's estimated *that just about 250 hours of HD programming is currently available for airing*. (wsj_1386)

Verb cues are by far the most frequent type of attributional cue, in particular reporting verbs which refer to a linguistic action, such as 'say' and 'declare', and opinion verbs expressing a cognitive process such as 'believe' and 'worry'.

Although verbs are the most common attribution anchor, syntactic cues can be expressed by other grammatical elements: nouns functioning as introductory elements (Renzi et al., 1995) as in Ex. (8); adjectives (Ex. (9)); prepositions and prepositional groups such as ‘for’, ‘in the eyes of’ and ‘according to’; adverbials such as ‘reportedly’ (Ex. (10)) and ‘allegedly’. For the attribution of speech acts, punctuation can be the only cue expressed, as in Ex. (11).

(8) However, Mr. Moran added that the Japanese generally have a positive view of the U.S. bond market because of expectations *that the dollar will remain strong and interest rates will decline*. (wsj_1213)

(9) **I**’m sure *they’ll formulate a reform that will be a recipe for the GDR’s future as a separately identifiable state*... (wsj_1875)

(10) *Japan Air Lines, Lufthansa German Airlines and Air France* reportedly *plan to form an international air–freight company this year, a move that could further consolidate the industry*. (wsj_1850)

(11) **Mrs. Thatcher**: “*If it’s one against 48, I’m very sorry for the 48.*” (wsj_1053)

1.2.2 Attribution in the Literature: A Fragmented Picture

1.2.2.1 Discourse Relation

Attribution has been considered as a type of discourse relation and annotated as such in a number of discourse resources, in particular is the **RST Discourse Treebank** corpus (Carlson and Marcu, 2001). This resource is grounded in the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) that establishes relations between *nucleus* and *satellite* discourse units. Although RST does not consider attribution as a rhetorical relation, the RST Discourse Treebank includes it. Attribution is annotated at the intra-sentential level and only when it is introduced by an attribution verb that refers to a speech or cognitive act and that takes a clausal complement. If the clausal complement is infinitival, then source and content are annotated together as a single discourse unit.

Because of the scope of the annotation, only certain types of attributions are annotated in the RST. This can be misleading for studies trying to identify or analyse attribution as they can mistake the annotation limitations as evidence of the type of structures expressing the relation. For example, the RST annotations were used by Skadhauge and Hardt (2005) to prove the redundancy of including attribution relations in the RST corpus, claiming that they can be automatically inferred based on syntactic information alone. Their study proves that a large proportion, but not all, of the attributions annotated in the RST corpus involve an attributional verb and its sentential object. However, this should not be misinterpreted as a proof of attribution being a syntactic relation, since it is based on a corpus that is not representative of attribution. Other discourse studies following the RST framework (Pardo and Nunes, 2003; Pardo et al., 2004; Afantenos et al., 2012) have also considered attribution as a rhetorical relation.

Attribution was also assimilated to discourse in the **GraphBank** corpus (Wolf and Gibson, 2005). This corpus describes it as an asymmetrical or directed relation (satellite-source to nucleus-content), holding between separate discourse segments. The content needs to be one or more sentences or a complementizer phrase. If this is not the case, it is treated as a single segment together with its source. Many attributions are therefore not identified due to the annotation constraints, in particular the one requiring discourse segments not to overlap. However, attributions can be nested into one another or overlap with other discourse relations. Another constraint is that only a single relation can hold between two discourse segments. Consequently, only one relation will be annotated when an AR holds between segments also affected by another relation.

According to the approach to discourse adopted by the **Penn Discourse TreeBank** (PDTB) (Prasad et al., 2008), attribution is not a discourse relation. While ARs relate an abstract object to an entity, discourse relations hold between abstract objects. Nevertheless, ARs are included in the annotation of the PDTB (see Sec. 2.1.1) because of the effect attribution has on the reliability and structure of discourse relations.

1.2.2.2 Opinion Carrier

A portion of ARs has been addressed and annotated by studies dealing with 'subjectivity analysis'. A subset of ARs, namely opinions and beliefs, corresponds to part of the 'private states' at focus in the MPQA Opinion Corpus (Wiebe, 2002). Private states are opinions, beliefs, thoughts, feelings, emotions, goals, evaluations and judgements.

Private states can be expressed by means of expressive subjective elements, such as sentiment bearing words, but also structures in common with attribution, such as an explicit mention of an opinion or belief or a speech event. The annotation is intra-sentential and comprises the text anchor, the source (see Ex. (12)), the target and also some properties relative to the private state, e.g. intensity and polarity.

(12) “The U.S. fears a spill-over,”said Xirao-Nima. (Wiebe et al., 2005, p.9)

Text anchor: said

Source: writer, Xirao-Nima

Text anchor: fears

Source: writer, Xirao-Nima, U.S.

Subjectivity analysis studies depend on the identification of ARs for the retrieval of the source of private states. A private state might correspond to the cue of an AR or be a term or expression inside an AR content. Despite a strong overlap in scope, the approach is considerably different. While a private state is defined as “an experiencer holding an attitude, optionally toward an object” (Wiebe, 2002, p.4), attribution goes in the opposite direction. The object, which corresponds to the content, is not optional, but a fundamental element of the AR.

1.2.2.3 Reported Speech

Reported speech represents a particular type of ARs. Studies concerned with the attribution of speech acts, or quotations (see Sec. 2.2.1), usually consider attribution as composed by *quotation*–**speaker** pairs (Ex. (13)). The cue element connecting speaker and quotation and expressing the type of AR (e.g. assertion or belief) is not annotated. While the attribution of quotations implies that what is attributed is an assertion, the cue is still a relevant element as it can greatly affect the AR as it expresses the attitude the source holds towards the content (consider the difference between ‘say’ and ‘deny’).

(13) *“The employment report is going to be difficult to interpret,”* said **Michael Englund**, economist with MMS International, a unit of McGraw-Hill Inc., New York. (wsj_0627)

Some studies, such as Glass and Bangay (2007) and Pouliquen et al. (2007), identify the textual anchor which establishes the relation. However, this anchor is considered as a device helping the extraction of an AR and not as integral part of the relation itself. In particular, lists of speech verbs are pre-compiled and their grammatical subject used to retrieve the quotation speaker. The quotation itself is identified with the grammatical object of the speech verb. While reported speech represents a large and important subset of ARs, other types of ARs exist and should be taken into account (e.g. beliefs, opinions, intentions).

1.2.2.4 Factuality and Events

ARs also affect temporal references, and ‘reporting’ has been included as an event class in TimeML (Pustejovsky et al., 2003a), a framework for the annotation of events. Reporting events have been annotated in TimeBank (Pustejovsky et al., 2006), a corpus of events and temporal references. Accounting for the relation between the time the document was produced and that of the reporting event represents a challenge. ARs insert an additional point in time, namely that of the enunciation in case of an assertion or the temporal point where a belief or fact was factual. For example, ‘John thought it was a good idea’ reflects John’s belief at a past point in time. This belief might have changed at the point the article was written or the present time.

Attribution has also strong implications for the factuality of the events expressed in the attributed span. This motivates its partial inclusion in FactBank (Saurí and Pustejovsky, 2009) where the content span is not marked, but events contained in it (e.g. ‘left’ in Ex. (14)) are linked to their source by Source–Introducing Predicates (SIPs) in order to derive their factuality. The SIP in Ex. (14) implies that the event underlined is considered by the source as just a possibility. The factuality of the event is assessed with respect to the source’s perspective as opposed to it being a fact of the world.

(14) Berven **suspects** that Freidin left the country in June. (Saurí and Pustejovsky, 2009, p.236)

1.2.3 Attribution and Discourse

1.2.3.1 Attribution Is Intertwined with Discourse

Attribution relations are closely tied to discourse relations as their inclusion in several discourse studies, presented in Sec. 1.2.2.1, shows. ARs have been variously included as a discourse relation itself (Wolf and Gibson, 2005; Carlson and Marcu, 2001) or as an attribute of discourse relations (Prasad et al., 2006). They were included in the PDTB since it was recognised that attribution affected polarity. Attribution also proved to be “a major source of the mismatches between syntax and discourse” (Dinesh et al., 2005, p.36).

If the arguments of a discourse connective are taken to be its syntactic arguments, attribution could lead to incorrect semantic interpretation. In Ex. (15) (Prasad et al., 2008, p.2966), the explicit discourse relation expressed by ‘while’ holds between the segments (15a) and (15c) and not between segment (15a) and the ‘purchasing agents’ saying what constitutes segment (15c). It is therefore important to recognise and exclude attribution from the discourse segment in such cases.

(15) a. Factory orders and construction outlays were largely flat in December (Arg1)

b. while (Conn.) **purchasing agents** said

c. *manufacturing shrank further in October* (Arg2). (wsj_0178)

While attribution is disruptive when annotating or recognizing discourse relations, the latter can benefit the annotation or recognition of ARs. Discourse relations may help the identification of *content* span boundaries, in particular for indirect ARs where the attributed span is not surrounded by quotation marks. Some studies (Skadhauge and Hardt, 2005; de La Clergerie et al., 2009) have taken an intra–sentential approach to attribution and restricted the AR content to being the grammatical object of a reporting verb. However, this is not a viable solution when dealing with a wider range of ARs. In some cases, discourse structure may play a role above the level of single sentences.

The ARs collected from the PDTB show that around 17% of ARs extend over more than one sentence (e.g. three sentences in Ex. (16)). Moreover, only half of these are attributions of direct quotations. English does not mark indirect reported speech grammatically, unlike German, where this is associated with subjunctive mood (Ruppenhofer et al., 2010). The resulting problem is how to determine the content span boundaries of indirect ARs when the syntactic structure would be of no help. While sometimes ambiguous also for human readers, recognising a content extending over more sentences could be in some cases achieved with the help of discourse relations.

(16) According to **Audit Bureau of Circulations**, *Time, the largest newsweekly, had average circulation of 4,393,237, a decrease of 7.3%. Newsweek's circulation for the first six months of 1989 was 3,288,453, flat from the same period last year. U.S. News' circulation in the same time was 2,303,328, down 2.6%.* (wsj_0012)

In Ex. (16), the last two sentences are a continuation of the content but they bear no syntactic relation with the first sentence. Instead, there are two discourse relations annotated in the PDTB and entailing an implicit connective *and* binding the first part of the content span with the second and the third sentence. Discourse alone might not provide sufficient evidence to determine the content extension. Nonetheless, in combination with other triggers such as verb tense and mood, this could help the correct identification of inter–sentential indirect ARs.

1.2.3.2 Attribution Is Distinct from Discourse Relations

The PDTB is rich in attribution annotation and represents a valuable starting point for the collection of a large resource for the study of attribution. However, what is annotated is not ARs but the attribution of discourse connectives and their arguments. Attribution is therefore subordinate to discourse and reconstructing a full AR can be rather complex.

The content of an AR might not fully correspond to a discourse relation or one of its arguments, but be composed of several discourse connectives and their arguments. We can consider the AR that corresponds to the second paragraph of Ex. (17): [4]

(17) The reports, attributed to the Colombian minister of economic development, said Brazil would give up 500,000 bags of its quota and Colombia 200,000 bags, the analyst said.

(HOWEVER) *These reports were later denied by a high Brazilian official, who said Brazil wasn't involved in any coffee discussions on quotas*, **the analyst said**.

(BUT) The Colombian minister was said to have referred to a letter that he said President Bush sent to Colombian President Virgilio Barco, and in which President Bush said it was possible to overcome obstacles to a new agreement. (wsj_0437)

The content span of this AR, is partially included in all three discourse relations below: the two implicit ones, having *however* and *but* as connectives, and the one with discourse connective *later*. In order to reconstruct the full AR from the annotation, it is necessary to take all three discourse relations into account and merge together the text spans attributed to 'the analyst said'.

1. The reports said Brazil would give up 500,000 bags of its quota and Colombia 200,000 bags (Arg1)

   HOWEVER (Implicit connective)

   These reports were later denied by a high Brazilian official (Arg2)

2. The reports said Brazil would give up 500,000 bags of its quota and Colombia 200,000 bags (Arg1)

   LATER (Connective)

   These reports were denied by a high Brazilian official (Arg2)

[4] Attribution annotations in the PDTB do not distinguish between source, cue and other circumstantial information. These elements are all part of the 'attribution span' which is indicated in bold in the examples.

3. *who said Brazil wasn't involved in any coffee discussions on quotas* (Arg1)

   BUT (Implicit connective)

   The Colombian minister was said to have referred to a letter that he said President Bush sent to Colombian President Virgilio Barco, and in which President Bush said it was possible to overcome obstacles to a new agreement (Arg2)

The example shows that there is no exact correspondence between ARs and discourse arguments and therefore some ARs are incompletely annotated or not annotated at all. This situation occurs when part of the AR content does not correspond to a discourse argument or when the whole AR is included in a discourse argument as in Arg1 of *But* (relation 3 above). The AR embedded in Arg1 ('**who said** *Brazil wasn't involved in any coffee discussions on quotas*') is just not annotated.

While the PDTB comprises several types of ARs that can be mostly reconstructed from the annotation, attribution still represents a distinct relation. This should be independently annotated since there is no exact correspondence between ARs and discourse relations. I therefore adopt an approach that separates the annotation of discourse and attribution.

1.3 Terminology

Attribution has been defined as "a relation of 'ownership'between abstract objects and individuals or agents" (Prasad et al., 2007, p. 40), with Abstract Objects (AO) referring to propositions, events or states. However, the object of an AR does not necessarily exactly overlap with an AO and can comprise parts of different AOs or be included in a single one. Carlson and Marcu (2001) state that speech acts and cognitive predicates should be marked as attribution, thus identifying attribution by the elements signalling them. However, verbs are not the only lexical anchor of ARs. Another definition is given by Murphy (2005, p. 131) who sees attribution as "the transferral of responsibility for what is being said to a third party". Although it considers only speech acts, this definition effectively captures the nature of attribution, a relation embedding different voices into one another.

The presence of a third-party, however, is not always necessary in order to have attribution, as in the embedded AR in Ex. (18). Attributions to the author or the same source as the attribution including them could be considered *pseudo-attributions*. These are redundant and are usually a device to present what is stated as more personal as opposed to factual. Their function is not that of linking linguistic material to its owner in order to transfer responsibility to them.

(18) But, says the general manager of a network affiliate in the Midwest, "**I** think *if I tell them I need more time, they'll take 'Cosby'across the street*." (wsj_0060)

The working definition of attribution adopted for the current study is that of a "relation ascribing the ownership of an attitude towards some linguistic material, i.e. the text itself, a portion of it or its semantic content, to an entity" (Pareti and Prodanof, 2010, p. 3566). This definition allows to consider both the attribution of opinions and that of direct and indirect speech acts without posing a limit on the nature of the attributed material, the entity or the text anchor. In particular, with 'linguistic material' we intend to capture both the actual span of text attributed and the semantic unit or units that correspond to a particular span of text. For direct attributions of speech acts, the ownership of the exact span of text is ascribed to the source entity. This span can be as little as a single word as well as sentences or paragraphs or even the complete text (in which case the source is its author). While it is also the semantic value of the text span that is attributed, it is relevant to note that also the exact words used to convey it are attributed. In all other cases, however, what is attributed is not the text span itself, but rather the semantic units it expresses, whether propositions, facts or eventualities.

Quotations represent a specific type of ARs and have attracted particular attention from the literature. Some studies only address direct quotations. Alrahabi et al. (2010, p. 162) defines them as "any kind of speech delimited by meta-characters (the typographical signs of quotation) and introduced by, at least, one linguistic marker referring to an act of speaking, whether the speaker is explicitly defined or not". However, quotations include also indirect and mixed (i.e. partly direct and partly indirect) reported speech. In this thesis *reported speech* and *quotations* will be used to refer to attributions of speech acts. Following the attribution type distinction adopted in the PDTB (Prasad et al., 2006), these will be generally called *assertion attributions*, although they are not limited to asserted statements.

Different terminology can be found also concerning the elements that constitute the AR. In the literature, the entity the material is attributed to is variously called. In quotation studies the source is usually called the *speaker* (Alrahabi et al., 2010), while opinion and subjectivity literature refers to it as the *opinion holder* (Kim and Hovy, 2005; Kim et al., 2007). Finally, some studies refer to it as the **source** (Choi et al., 2005; Prasad et al., 2008; Carlson and Marcu, 2001; Wiebe et al., 2005). For the current work, the latter term will be adopted, since both opinion and assertions will be considered for attribution. When referring only to quotation or to opinion attributions, *speaker* and *opinion holder* are also used.

The linguistic material attributed to the source can be of different nature and thus has been named as *quote* or *quotation*, *reported speech*, *propositional opinion* (Bethard et al., 2004) and *inside* (Wiebe et al., 2005). In this thesis, I will refer to it as the *content* following the terminology adopted by Bergler (1992).

The lexical anchor signalling the attribution is a key element of this relation. Some annotation projects have merged the cue with the source and annotated both elements together as the *reporting span* or the *attribution phrase* (Prasad et al., 2008). Others, such as Pouliquen et al. (2007), do not annotate the cue although they make reference to *reporting verbs* and use them to identify the speaker of a quotation. I will refer to it as the **cue** as this reflects its function of 'attribution signal' and is suitable for different types of cues, such as those expressing opinions, intentions or assertions.

In addition to source, cue and content, some supplementary information can also be contextually inserted. As discussed in Pareti (2009), this includes circumstantial elements, such as a specification of the event as in Ex. (19)[5], as well as topic or audience of the original utterance. Although not strictly necessary, supplement information may provide key elements for understanding the attribution itself. I will refer to it as *supplement*.

(19) IN AN INTERVIEW AT THE TIME OF HIS RESIGNATION FROM MCA, **he** <u>said</u>: "*I'd rather build a company than run one.*"(wsj_0408)

Some studies (Kim and Hovy, 2006b; Lu, 2010) consider also the *target* of an attribution, namely the topic or the entity the statement or opinion is about. Although relevant to fully comprehend the content, the target of an AR does not affect the relation itself and is therefore not included in the scope of this work.

1.4 Attribution Extraction: Task Definition

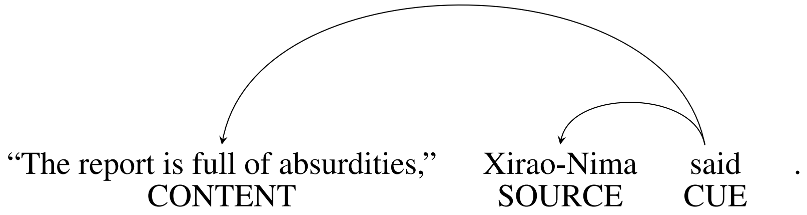
In this thesis, I adopt a lexically anchored approach to attribution, i.e. source and content are identified by the text spans expressing them and also the relation itself is lexically anchored to its cue element (Ex. (20)).[6] I propose to recast AR extraction as

[5] Supplement elements are indicated in the examples in small capital letters.

[6] This sentence is taken from an example in Wiebe et al. (2005, p.11).

the task of identifying and linking the text span expressing source, cue and content elements of each AR.

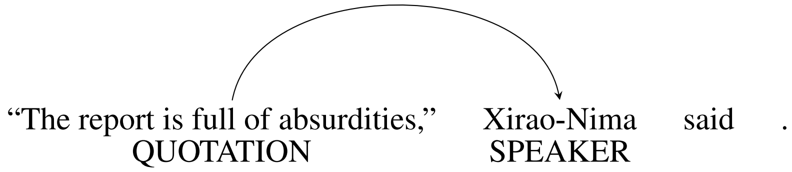
(20)



“The report is full of absurdities,”	Xirao-Nima	said	.
CONTENT	SOURCE	CUE	

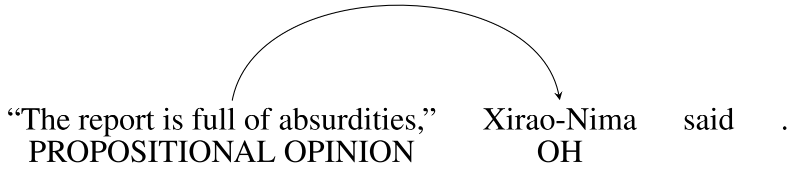
This approach differs from that of traditional Speaker Attribution literature (e.g. Pouliquen et al. (2007); Elson and McKeown (2010)) whose task is to link quotation content spans to the entity who uttered them. Thus, while the quotation content is a span, the source is an entity and no cue element is identified (Ex. (21)). The approach also differs from Opinion Analysis studies. Some studies, such as Bethard et al. (2004), identify the content expressing a propositional opinion and link it to its source or Opinion Holder (OH)) (Ex. (22)), similarly to Speaker Attribution studies. Others (e.g. Wiebe (2002); Wiegand and Klakow (2010)) identify opinion or emotion expressions, similarly to the cue, and link them to their source without identifying the span the opinion refers to, but just the opinion target (Ex. (23)).

(21)



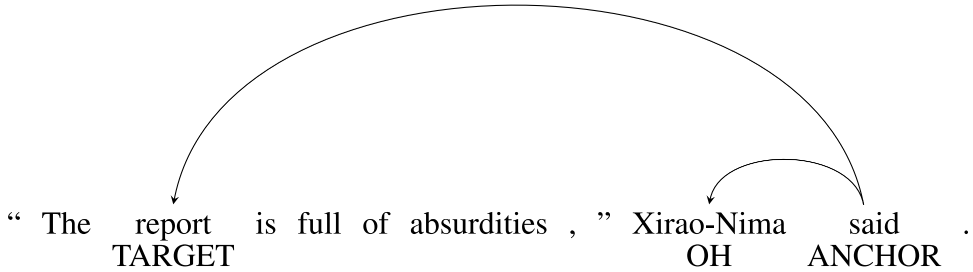
“The report is full of absurdities,”	Xirao-Nima	said	.
QUOTATION	SPEAKER		

(22)



“The report is full of absurdities,”	Xirao-Nima	said	.
PROPOSITIONAL OPINION	OH		

(23)



“	The	report	is	full	of	absurdities	,	”	Xirao-Nima	said	.
		TARGET							OH	ANCHOR	

1.5 Contributions

The main claims of this dissertation are that: (1) attribution is best annotated independently from other linguistic phenomena and it is possible to annotate different types of attribution with sufficient agreement; (2) attribution relations can be viewed as lexicalised and their extraction is best addressed as the identification of the spans expressing their source, cue and content elements; (3) having a large annotated resource, supervised computational models can extract complete attribution relations with better than state of the art results. In support of these claims, this dissertation makes the following contributions:

- Creation of the first large-scale attribution corpus, comprising a wide range of ARs types and structures:
    - Adaptation of a scheme to annotate relations of attribution in text.
    - Inter-annotator agreement study to evaluate the scheme applicability.
    - Collection and further semi-automatic annotation of a preliminary corpus of around 10k ARs.
    - Semi-automatic identification and manual annotation of an additional 10k ARs not annotated in the preliminary corpus.
    - Release of the complete resource.
- Definition of a methodology for the automatic extraction of attribution relations:
    - Creation of a classifier for the automatic identification of verb-cues.
    - Creation of a model for the automatic labelling of content spans.
    - Extension of a speaker attribution system.
    - Definition of algorithms for recovering full cue and source spans and for matching the cue to its source and content.
- Preliminary exploration of ARs in other genres and of the ways attribution features can be identified and employed to select and present information.

1.6 Outline

The remainder of this thesis is organised as follows. Ch. 2 presents an overview of related literature. Studies originating from different research areas whose scope overlaps with attribution are also included. This will show the scarcity and limitation of current available resources and motivate the need to create a large corpus of ARs. It will also present the state of the art concerning the automatic extraction of attribution, with many studies addressing only a subset of it and without achieving high accuracy. The applications based on those extraction systems are also presented.

Ch. 3 describes the development of a large attribution corpus. The adopted schema is presented and tested by conducting an inter-annotator agreement study. The collection and further annotation of the corpus is described and the final resource is used to explore and present the many ways ARs are encoded.

Ch. 4 proposes a methodology for the automatic extraction of ARs. This makes use of the newly created corpus to train and test supervised models to extract the spans expressing each of the AR components: source, cue and content. The models are compared to heuristics models similar to those adopted in the literature showing significant improvements.

Ch. 5 will discuss the portability of the proposed methodology for attribution annotation to other languages and other genres. In particular, I will compare the encoding of attribution in Italian vs English and describe some preliminary joint work on attribution in speech.

Ch. 6 presents some potential future directions of the present work. The extraction of the attribution components could be modelled as a joint problem. Concerning potential applications of AR extraction, I propose to make use of ARs to select information based on properties of the AR and its source as well as to enhance news summarization.

Finally, Ch. 7 will sum up the key findings and contributions of this thesis.

Some of the work presented here was previously published[7] and presented at conferences. Specifically, parts of of Ch. 1 were published in Pareti (2012b), parts of Ch. 3 in Pareti (2011, 2012a), parts of Ch. 4 in Pareti et al. (2013) and parts of Ch. 5 in Pareti (2015).

[7] Publications available from: `http://homepages.inf.ed.ac.uk/s1052974/`

Chapter 2

Related Work[1]

Attribution relations have been annotated as discourse relations, attributes of discourse relations, structures carrying factuality, frames for the expression of subjective language, quote–speaker relations and classes of temporal references. While this proves their relevance for different domains, whether as disruptive elements to rule out or essential carriers to identify, this fragmented effort has produced only a limited and marginal picture of this relation.

In Sec. 2.1, I will review existing resources annotated with some aspects of attribution and highlight their limitations, particularly in terms of limited size or only partial coverage of attribution, that motivate the effort of creating a new large corpus.

Sec. 2.2 will present a structured review of attribution extraction studies to date and discuss the approaches adopted highlighting their limitations in terms of scope and assumptions and the relatively low results they report. Due to the lack of large and complete annotated resources, extraction studies have often resorted to rule-based systems that are not adequate to identify a wide range of attribution structures.

The applications available to date, based on attribution, will be briefly presented in Sec. 2.3. This will give an idea of the potential and relevance of attribution sensitive tools, particularly for opinion mining and information extraction tasks, but also show their current limitations, responsible for the still limited uses of these tools.

2.1 Corpora

This section reviews existing resources annotated with ARs or some aspects thereof. The corpora are grouped into: news corpora, narrative corpora and corpora in lan-

[1] Part of this chapter was published in Pareti (2012b)).

guages other than English.

2.1.1 News Genre

Penn Discourse Treebank (PDTB)

The PDTB 2.0 (Prasad et al., 2008) is a collection of over 2,000 news articles from the Wall Street Journal (WSJ) annotated with discourse connectives and their arguments. Attribution is not annotated as a discourse relation itself but just when it overlaps with discourse relations (see Sec. 1.2.2.1). Both discourse connective and its arguments are associated with their source and some attributes.

The source can be annotated as: *writer*, i.e. there is no AR in the text and therefore the default connection of the article to its writer is assumed; *other*, in case of an attribution to a source other than the writer; *arbitrary*, if the source is not a specific one (e.g. ‘one’, ‘some’) or not mentioned (e.g. passive forms, ‘reportedly’). In addition to the source, the PDTB annotates the attribution type which reflects whether the content is presented as a fact (introduced by e.g. ‘know’, ‘hear’, ‘remember’), an assertion (e.g. ‘say’, ‘whisper’), an eventuality (e.g. ‘order’, ‘want’) or a belief (e.g. ‘think’, ‘doubt’).

Finally, two additional features, *determinacy* and *scopal polarity*, account for the factuality of the attribution itself, i.e. whether the relation between source and content is presented as a fact of the real world or an unreal or hypothetical fact. The factuality of what the content expresses is not evaluated. An AR such as ‘John could say that the earth is round’ would be non factual since the attribution of the content to the source is only hypothetical and not presented as a fact.

Since attribution relations have only been annotated with respect to discourse connectives (both explicit and implicit) and their arguments, there are several places where attribution is not annotated. Attribution is not annotated when there is no discourse relation and only an entity-based relation of coherence is marked (labelled as EntRel) and when two adjacent sentences are not joined by a discourse or entity-based relation (labelled as NoRel). This is also the case when the AR does not participate in an annotated discourse relation as in Ex. (24). Since discourse relations are only annotated across paragraphs when there is an explicit discourse connective, in the example the AR is not annotated since it constitutes the entire paragraph. ARs are also not annotated when their *attribution span* (i.e. the span corresponding to source and cue) is itself part of the argument of a discourse connective (e.g. Ex. (25)) and in case of nested attributions.

(24) *A form of asbestos once used to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago,* **researchers** reported. (PDTB 0003)[2]

(25) [The National Association of Manufacturers settled on the Hoosier capital of Indianapolis for its fall board meeting]$_{\text{Arg1}}$. [And]$_{\text{Connective}}$ [**the city** decided *to treat its guests more like royalty or rock stars than factory owners*]$_{\text{Arg2}}$. (PDTB 0010)

In addition, parts of an attribution content outside the discourse connective arguments span are not included in the annotation as in Ex. (26). Non-clausal attribution spans such as 'according to' are included in the argument span they attribute. In the PDTB 2.0 there is no distinction between source and cue as they are annotated together in the attribution span. The attribution may also include additional circumstantial information, e.g. *the judge quipped* IN AN INTERVIEW (wsj_0049) or *The U.S. government* IN RECENT YEARS *has accused* (wsj_0051).

(26) *[The asbestos fiber, crocidolite, is unusually resilient]$_{Arg1}$ [once]$_{Connective}$ [it enters the lungs]$_{Arg2}$, with even brief exposures to it causing symptoms that show up decades later,* **researchers** said.(PDTB 0003)

RST

The RST Discourse Treebank (Carlson and Marcu, 2001) consists of 385 news articles from the WSJ, drawn from the Penn Treebank (Marcus et al., 1993). A relation of attribution is established between a nucleus, i.e. the content, and its satellite, i.e. the source (Ex. (27)). In this resource however, attribution is annotated only at the intra–sentential level and requires the existence of an explicit source. Only attributing verbs and the expression 'according to' are considered possible attribution signals, while other cues are not taken into account.

(27) [The impact won't be that great,]$_{\text{Nucleus}}$ [said Graeme Lidgerwood of First Boston Corp.]$_{\text{Satellite}}$ (wsj_1111)

The corpus was annotated by 36 professional language analysts with annotation experience, fully-trained, using an extensive annotation manual. The raw agreement among 6 annotators is reported on 3 tasks: the detection of the annotation span, the

[2] The code refers to the PDTB texts the examples are taken from.

nuclearity and the choice of relation. Their agreement on the latter, on documents already segmented for discourse units, was .71.

GraphBank

Another resource including partial annotation of attribution, considered as a discourse relation, is the GraphBank (Wolf and Gibson, 2005). This consists of 135 texts from the WSJ and Associated Press Newswire. Two students annotators were trained and provided with an annotation manual. They achieved an agreement of 0.84 (both Kappa and raw agreement) for the discourse segment grouping task and Kappa agreement of .83 for relation selection. In Ex. (28), the relations are established as follows (Wolf and Gibson, 2005, Table 5, p.269):

- *Elaboration* between 1a and 1b
- *Same* between 1 (or 1a) and 4
- *Attribution* between 2 and 3
- *Elaboration* between 2-3 and 1 (or 1a and 1b)
- *Attribution* between 4 and 5
- *Violated expectation* between 2-3 and 4-5

(28) [Mr. Bakers assistant for inter-American affairs,]$_{1a}$ [Bernard Aronson,]$_{1a}$ [while maintaining]$_{2}$ [that the Sandinistas had also broken the cease-fire,]$_{3}$ [acknowledged:]$_{4}$ [“Its never very clear who starts what.”]$_{5}$ (wsj_0655)

MPQA Opinion Corpus

The Multi-Perspective Question Answering (MPQA) (Wiebe, 2002) opinion corpus consists of 692 documents from different U.S. and non-U.S. news sources such as the WSJ and the American National Corpus (ANC). The attributions included in the annotation are those introducing the so called ‘privates states’, namely opinions, beliefs, thoughts, feelings, emotions, goals, evaluations and judgements. The annotation comprises: explicit mentions of private states, speech acts introducing a private state and expressive subjective elements, i.e. words and expressions indirectly conveying a private state such as ‘absurd’.

The annotation is limited to the intra-sentential level and distinguishes three elements: the *text anchor*, the *source* and the *target*, i.e. what the opinion or attitude refers to. Some properties relative to the private state such as intensity and polarity are also annotated. Sources can be nested, that is, not only may the direct source of a private state or speech act be annotated, but also any source reporting the private state, up to the writer of the text itself.

This is an important aspect, since statements and opinions are attributed and filtered by the additional sources of the text including the statement or opinion. Being able to identify all sources allows better judgement of the factuality of the attribution and the information it conveys. It enables taking expertise and bias of each source into account. Most attribution studies, however, do not consider sources other than the direct source of a statement or an opinion.

In the Ex. (29) (Wiebe et al., 2005, p.11) the AR cue 'said' is annotated as a speech act introducing a private state. Its sources are the AR source as well as the writer of the text. The target is 'report', that is the element the private states 'full of absurdities' refers to.

(29) "The report is full of absurdities," Xirao-Nima said.

Speech act introducing a private state:
Text anchor: said
Source: writer, Xirao-Nima
Target: report

Expressive subjective element:
Text anchor: full of absurdities
Source: writer, Xirao-Nima

A portion of 13 articles was annotated by three non–expert annotators, following extensive training. They were provided a manual containing a case study and the general idea behind the annotation and a separate document with the specific annotation instructions based on examples, with particular stress on the importance of context. Since the annotators would annotate different expressions to identify text anchors, they propose an alternative to Cohen's Kappa agreement, the *agr* metric. This is a directional measure defined as $agr(a||b) = |A\ matching\ B|/|A|$, where a and b are a pair of annotators and '*A matching B*'comprise the elements annotated by a that were also annotated by b. With this measure, they report an overall agreement of .82 for the identification of explicitly mentioned private states and speech events. These categories partly correspond to attributions categories also in the scope of this thesis.

Although ARs are partly annotated in the MPQA Opinion Corpus, the annotation is incomplete. Only ARs expressing or containing private states are annotated. The annotation includes the cue (textual anchor) and source(s) of an AR, but not its content span. This resource has inspired several studies and was employed to model and test the extraction of opinion ARs (Sec. 2.2.3).

NTCIR-7 corpus

A set of corpora annotated with opinions and their polarity as well as their holder and target was developed for the NTCIR-6, 7 and 8 Multilingual Opinion Analysis Task (MOAT) (Evans et al., 2007; Seki et al., 2008, 2010). The corpora in NTCIR-7 and 8 comprise documents sampled from different topics selected from news in English, Japanese, Simplified and Traditional Chinese. Similarly to the Opinion Corpus (Wiebe, 2002) opinion holders are assigned an unique ID. Non-opinionated attributions are not considered nor are attributions to anonymous sources. Attributions to the author are annotated. In addition, if different clauses in a sentence have different opinion holders, the opinion holder of the main clause is assigned as the source of the whole sentence. In Ex. (30), the clausal text bearing an attitude, in this case neutral, is associated with its target as well as the opinion holder. The corpora differ in size, from around 150 documents to almost 800 and in opinion clauses that are annotated, from approximately 4.5k to 9.5k.

(30) [Ji Man-ho]$_{\text{OpinionHolder}}$, publisher of the monthly magazine and president of Maeil Health Magazine Co., said, ["In the 21st century, [Korean traditional medicine]$_{\text{Target}}$, while improving people's health, also needs to make a great effort to re-examine its role as an independent medical science."]$_{\text{Attitude}}$ (NTCIR-7 KT2001_03549)

TimeBank

The corpus (Pustejovsky et al., 2003b) consists of 183 articles from the WSJ, New York Times (NYT), Associated Press (AP) and transcribed news reports and is annotated with events and temporal expressions. Attribution overlaps mainly with the events labelled as: REPORTING, PERCEPTION, I_STATE and I_ACTION. Subordinating links (SLINK) provide the connection between the attribution-bearing event (e.g. 'said', 'reports') and the event(s) in its complement span. In Ex. (31), the Event Selecting Predicate (ESP) 'said', from the class REPORTING, is linked to the event 'infatuated' via an SLINK of type evidential.

(31) Newspaper reports have [said]$_{\text{ESP}}$ Amir was [infatuated]$_{\text{Event}}$ with Har-Shefi [...]

FactBank

FactBank (Saurí and Pustejovsky, 2009) is a superset of TimeBank, including all of its texts with the addition of some other ones. Both corpora are annotated with time and event expressions following the TimeML framework (Pustejovsky et al., 2003a). Overall, the corpus comprises 9,488 events, each associated with one or more sources and factuality pairs. All events inside an AR content are associated with the AR source and all above sources, including the author of the text. Source candidates were automatically extracted by taking all NP heads in a certain syntactic relation with the Source Introducing Predicate (SIP). This rules out all sources that are not syntactically related to the AR cue. Sources that were not among the available candidates were manually added at a later stage. SIP were identified with .88 and sources with .95 Kappa agreement.

With respect to TimeBank, FactBank also annotates the source(s) of the event and its factuality value. In Ex. (31), the event 'said' would be attributed to the author, while 'infatuated' to both the author and 'reports'. Both are associated with a factuality value CT+ which means the source is certain about the event.

2.1.2 Narrative Genre

Narrative is an extremely complex genre for ARs, since a wide range of structures can be used to express attribution, with great style differences. While direct quotations in news are more strictly encoded, usually using double quotation marks, in narrative these can be replaced for example by single or double angle brackets (<...>), spacing or dashes as in Ex. (32). Moreover, the text can have a dialogical form, close to transcribed spoken language, where turns are not necessarily lexically marked since they can be inferred by other means (e.g. the alternation of turns or a certain choice of words that characterises a speaker). In Ex. (32), "Yes, my love?" is not explicitly attributed to the character called Mulligan. The source can be inferred because the previous dialogue turn is addressing it. The following turn is also not explicitly attributed, since the dialogue alternation makes it clear this is Stephen's turn. This flexibility is not only due to narrative being a creative style, but also to the limited pool of characters that are available at a given point and the provided context that can help disambiguate among potential sources.

(32) Buck Mulligan's gay voice went on.

> My name is absurd too: Malachi Mulligan, two dactyls. But it has a Hellenic ring, hasn't it? Tripping and sunny like the buck himself. We must go to Athens. Will you come if I can get the aunt to fork out twenty quid?
>
> He laid the brush aside and, laughing with delight, cried:
>
> -Will he come? The jeune jesuit!
>
> Ceasing, he began to shave with care.
>
> -Tell me, Mulligan, Stephen said quietly.
>
> -Yes, my love?
>
> -How long is Haines going to stay in this tower?
>
> (James Joyce's Ulysses, Episode 1 (Joyce, 2001, p. 4))

There is only one corpus of narrative texts annotated with attribution in English: the **Columbia Quoted Speech Attribution Corpus** (CQSA) (Elson and McKeown, 2010). The corpus is only annotated with direct quotations, thus missing a wide range of ARs and the structures expressing them. It comprises excerpts from 11 narrative works from the 19th and 20th century written by 6 different authors. Overall 3,578 direct quotations and their speakers are annotated. The annotation was performed through Amazon's Mechanical Turk. Three annotators were asked to link each quotation to its speaker and cases were there was no majority agreement were discarded. NEs as well as nominal mentions were automatically tagged.

2.1.3 Languages Other than English

The literature has produced a small number of corpora annotated with ARs in languages other than English. These resources are limited in size which makes them poor candidates for training supervised attribution extraction systems, but they are nonetheless worth mentioning. They originate from the annotation of discourse relations, opinion frames or quotations.

Italian Attribution Corpus (ItAC), Italian

This corpus is part of preliminary work (Pareti, 2009; Pareti and Prodanof, 2010) I previously developed for the annotation of attribution relations in Italian news texts. The goal was to develop an annotation scheme for attribution in Italian, able to comprise several aspects and issues of this particular relation. ItAC[3] is built from a balanced

[3] Available at: `http://homepages.inf.ed.ac.uk/s1052974/resources.php`

TAGS	ATTRIBUTES
attribution role	*content, cue, source, supplement*
type	*assertion, belief, fact, eventuality*
source	*writer, other, arbitrary, mixed*
factuality	*factual, non factual*
scopal change	*none, scopal change*
relation	*set_n*

Table 2.1: ItAC Annotation Scheme. (Pareti and Prodanof, 2010)

selection of 50 articles (37k tokens) from the Italian Syntactic Semantic Treebank corpus of newspaper articles (ISST) (Montemagni et al., 2003). It comprises texts from major Italian newspapers (i.e. Il Corriere della Sera, La Repubblica and Il Sole 24 Ore) published between 1985 and 1995. The corpus was annotated using the MMAX2[4] annotation tool (Müller and Strube, 2006), an open-source tool that allows overlapping and discontinuous annotations.

There are 461 ARs in the corpus. The annotation scheme summarized in Table 2.1 was inspired by the PDTB annotation of attribution and comprises all the types of ARs and the features described in the PDTB. The annotation is extended in order to further annotate the attribution span by separately identifying the span expressing the source, the one expressing the attitude (i.e. the cue, e.g. a belief, an order) and possible circumstantial information (supplement). Each AR component is assigned its attribution role and associated with the other components in the same relation. The features were annotated on the cue element.

In the annotation scheme there are no constraints on the kind of cue introducing the attribution nor on what is considered as content, e.g. a single word, a phrase, several sentences. Moreover, the scheme allows for discontinuous contents and nested attribution to be annotated.

German Political News Opinion Corpus (DE News), German

This corpus (Li et al., 2012) annotates opinions and sentiments in German political news, inspired by the MPQA corpus (Wiebe et al., 2005). The 108 documents in the corpus are double annotated with the *source* (i.e. the attitude holder), the *target* (i.e. who or what the attitude is about), the *text anchor* (i.e. the text span expressing the

[4] Available open-source from http://mmax2.sourceforge.net/

attitude) and the *auxiliary* (i.e. words affecting the attitude such as negations and intensifiers) of an opinion. Three opinion related features are also included: the attitude, its intensity and whether the attitude is context-dependent, i.e. it derives from other elements in the context. The annotators identified 315 opinion frames. Unlike the MPQA annotation, the focus is here only on the identification of sentiments associated to words and not on the identification of all private states. Thus quotations and other types of attributions are not identified.

CorpusTCC and RHETALHO, Brazilian Portuguese

Part of a project for building a discourse parser, the first corpus (Pardo and Nunes, 2003) consists of 100 scientific texts (about 53,000 tokens) from the computer science domain, while a second corpus (Pardo et al., 2004) comprises 50 scientific and online news texts (from Folha de São Paulo). Both corpora[5] are annotated following the RST annotation tagset and protocol (Carlson and Marcu, 2001). CorpusTCC has 185 relations labelled as 'attribution'.

GloboQuotes, Portuguese

The corpus (Fernandes et al., 2011) is a collection of 685 news texts, from 10 news genres, published between 2007 and 2008 on the `globo.com` portal. It is annotated with quotations and their speakers, NEs and coreference. In spite of its size, the corpus contains only 1007 quotations. This might be a result of a lower incidence of quotations in certain news genres included in the corpus or derive from some constraints on the annotation.

ANNODIS, French

ANNODIS (Afantenos et al., 2012) is a resource consisting of 156 texts (i.e. news, Wikipedia, research and reports) and around 687k tokens. The resource annotates discourse structures (rhetorical relations and multi-level structures). Attribution is among the rhetorical relations annotated. However, given that attribution is annotated when other rhetorical relations are not also identified, there are only 75 instances of attribution in the corpus.

2.1.4 Discussion

While several resources comprise some annotation of attribution, these resources are mostly too small or incomplete to be employed to train supervised extraction systems.

[5] Available at: `http://www.icmc.usp.br/~taspardo/Projects.htm`

A comparison of the most relevant resources is presented in Table 2.2. This shows that most resources consist of English news texts. Apart from quotation corpora, attribution is usually not directly annotated, but included as a discourse relation or opinion expression. Corpora can be as small as 40 texts and contain only few attributions.

Corpus	Annotations	Texts	Genre	Language	Type
PDTB	10k	2,159	news	EN	discourse, ARs
RST	small	385	news	EN	discourse
GraphBank	small	135	news	EN	discourse
MPQA	-	692	news	EN	opinions
NTCIR	4.5k-9.5k	150-800	news	EN/JA/ZH	opinions
TimeBank	small	183	news	EN	events
CQSA	3.5k	11 books	narrative	EN	quotes
ItAC	461	50	news	IT	ARs
DENews	315	108	news	DE	opinions
CorpusTCC	185	100	scientific	PT	discourse
RHETALHO	small	40	various	PT	discourse
Annodis	75	156	various	FR	discourse
GloboQuotes	1007	685	news	PT	quotes

Table 2.2: Overview of relevant resources annotated with attribution relations, a portion of it or other relations overlapping with attribution.

The only relatively large resources are the PDTB, the MPQA, the NTCIR and the CQSA corpora. None of them is fully annotated with ARs. In the MPQA and NTCIR corpora, attribution is partly annotated, together with opinions and sentiments. While the annotation of cue and sources is included, the text span corresponding to the content is not annotated. The CQSA instead annotates direct quotations only and does not comprise the annotation of the AR cue. In the PDTB, discourse connectives and arguments are potential AR contents for which an attribution span including source and cue mention is usually annotated. Attributions are missing or incomplete when not fully matching explicit discourse relations.

Since no available resource is fully satisfactory, part of the goal of this thesis was to create a large and complete resource able to support a wide range of studies and the development of automatic extraction systems. While incomplete, the PDTB was chosen as the starting point to develop such attribution corpus, since it comprises a large number of attributions and the annotation is more compatible with the proposed approach to attribution. The corpus development is described in details in Ch. 3.

2.2 Attribution Extraction

Recognising the source of a piece of information or an opinion is of great importance for several tasks and could have useful applications, particularly in information extraction, multi-perspective question answering and opinion analysis. Although several studies have developed strategies for the automatic extraction of attributions, only a subset of the possible ARs has been taken into consideration. A comparison of their result is particularly complex since these studies, although addressing the same issue, take different perspectives.

Not only have attribution extraction systems been developed for different languages and different domains, but they have been tested on incomparable test sets and report results only on the portion of ARs they aim to recognise. It is therefore necessary to compare the scores they present by taking the portion of attribution they consider and their evaluation method into account.

This section will present a classification of attribution extraction studies to date and of the approaches adopted. A preliminary distinction can be made between approaches focusing on reported speech only (Sec. 2.2.1), mostly addressing the identification of the source of a given quotation, i.e. the *speaker*, and approaches in the field of Opinion Analysis (Sec. 2.2.3), primarily addressing the attribution of opinions to their *opinion holder*. Both areas have developed in recent years, starting from the work of Zhang et al. (2003), Bethard et al. (2004) and Choi et al. (2005), but follow rather distinct paths.

2.2.1 Quotation Studies

2.2.1.1 Direct Reported Speech

Within studies dealing with the automatic attribution of reported speech, it is also possible to identify distinct subgroups sharing different motivations and scope. First of all, while some studies have addressed both the attribution of direct and indirect reported speech, others have focused on direct reported speech exclusively, applied either to the narrative or news domain. In narrative, the interest for the source of a quotation originated from the necessity to identify the speaker character of each quotation in order to extract scripts and assign different voices for synthesising children's stories (Zhang et al., 2003; Mamede and Chaleira, 2004). One of the main challenges encountered by these studies is that of handling aliases of each character in order to recognise possible different mentions as the same voice in the narration. In addition, chains of quotations from the same speaker, with just the first one being attributed, are also common in narrative, typically in dialogues between characters.

The approach developed by Zhang et al. (2003) starts with identifying quotations, including nested ones, and determining whether they are new or a continuation of previous quotations. In that case, they inherit the same source assigned to the preceding quotation. Afterwards, they extract NEs from the story and consider only the ones belonging to its characters. Noun phrases are also taken into account for the identification of character names that are not proper nouns. Finally, each new quotation is assigned to a speaker chosen according to basic rules: a character mention in the same-paragraph, preceding the quote, or otherwise a named character following the quote. Characters' names that are proper nouns are also given higher priority and the proximity to reporting verbs is also considered. Their speaker attribution system achieved 47.6-86.7% accuracy, depending on the style and complexity of attribution in different stories.

Subsequently, Glass and Bangay (2007) performed the same task on novels, using a salience-based method to identify a speech verb near the quotation and its *actor*, which they assume points to the speaker. This is chosen from a pre-compiled list of characters participating in the scene. Verb salience is determined by assigning a score to each verb considering whether: it is a main verb; it has *communicate*, *verbalise* or *breathe* as a hypernym in WordNet (Fellbaum, 1998); it is in an adjacent sentence and its proximity to the quotation. Actor salience is computed considering: its distance from the verb and if it is the subject or object of the verb; if it is a noun having *person* as its hypernym in WordNet; if it is a pronoun or a foreign noun, a proper noun or a title. If the actor is an anaphoric reference, then the speaker is derived from the previous actor or the last speaker. The system is evaluated on a corpus of 13 novels and achieves an average accuracy of 81.71% for the identification of the speaker and 79.4% including the identification of the actors. This approach does not consider cues other than verbs.

A similar approach to the attribution of direct quotations in narrative is taken by Elson and McKeown (2010). It also derives possible reporting verbs from WordNet. Quotations are divided into categories reflecting the way the quotation is attributed to the speaker, such a quotation followed by a reporting verb and an entity or a quotation followed by a pronoun and a reporting verb. NEs are extracted using the Stanford NE tagger (Finkel et al., 2005). The system makes a binary classification for each candidate entity, up to 15, which also predicts the probability of a given entity to be the speaker based on the extracted feature vector for that particular quotation category. Different strategies are then proposed to reconcile the predictions. The system takes dialogue chains into account and makes use of gold-standard information to generate some of the features. The speaker attribution system, tested on the 3,578 quotations in the CQSA corpus (see Sec. 2.1.2), achieves 83% accuracy.

Inspired by the work by Elson and McKeown (2010), He et al. (2013) developed a speaker identification system for novels that can be applied to texts other than the training ones, therefore with a different set of speakers. Candidate speakers are identified using a rule-based procedure. This looks for a speech verb before, after or between quotes, starting from a pre-defined list of 12 verbs (i.e. say, speak, talk, ask, reply, answer, add, continue, go on, cry, sigh, think). If none of these verbs is found, then any verb preceded by a noun or personal pronoun is selected. The subject of one of these speech verbs, or the identified verbs, is taken as a speaker candidate. Only the two most immediate speakers in each direction are taken into account. The ranking system uses a set of features based on lexical and syntactic clues, including speaker alternation, eventual vocatives inside the quotation and the lexical context. Depending on the novel, the speaker attribution achieves 74 % to 82% accuracy.

Non-fictional quotations, particularly from well-known sources, have attracted considerable attention and manually annotated collections appear in several websites. Being able to automatically gather quotations has become a very relevant task in the news domain, which presents different characteristics and challenges with respect to narrative. One of the most prominent projects in this area is the one by Pouliquen et al. (2007) both because of its application in the News Explorer system[6] (Sec. 2.3), and because it is the only project aiming at the cross-lingual identification of sources of quotations.

The system collects thousands of news reports in 11 languages on a daily basis and is therefore based on an attribution strategy highly independent from deep linguistic analysis. A small list of reporting verbs and 6 lexical patterns specify regular expression to be matched against the text. Pronoun or full noun anaphoric sources are not handled and only full mentions of NEs are recognised. NEs are then stored in a database, together with all their name variants and some frequent collocations found

[6] Accessible at: http://emm.newsexplorer.eu/

in the different languages.

While on the one hand NewsExplorer is able to collect and present direct quotations in online news articles for several languages and seems to collect a great amount of them daily, on the other hand the system is able to recognise merely the tip of the iceberg of attribution relations. Pouliquen et al. (2007) reports results on a small test set of 55 quotations collected for evaluation purposes. They identified 76% as having a structure that their patterns could not identify. On the remaining 24% of quotations, they report a *recall* of 54% (which corresponds to a *recall* of 13% over all quotations in the test set). Out of the 8 quotations their system extracted, 7 were correctly identified.

Despite the limitation of such a small test set, from their results it is clear that while the use of fixed lexical patterns might achieve a relatively high precision, their recall power is extremely low since attribution relations present great internal variation within and across languages. Although Pouliquen et al. (2007) claim that the extremely high redundancy of the online news domain assures that in the end 100% of the quotations will be at least once extracted by their system, nonetheless relevant quotations might be missed or their identification delayed – for example, less popular sources mentioned in more local news or scoops that would have to wait to be echoed by other news providers before being identified. While NewsExplorer implements an innovative concept for exploring news, also across languages, the full potential of the project is still to be achieved. This would require a more flexible extraction system able to recognise several attribution structures, extend to indirect and mixed quotations and retrieve anaphoric or common noun source mentions.

The last study addressing the attribution of direct reported speech only is by Liang et al. (2010) and was used by the Evri search tool (Sec. 2.3). Similarly to the previous study, thousands of quotations from news articles and blogs are collected daily, the strategy adopted being also that of using a list of reporting verbs and some patterns. The role of source is simply assigned to the subject of the reporting verb. The systems performs more sophisticated linguistic analysis than NewsExplorer since it processes each text with sentence splitting and parsing and also performs coreference resolution to link anaphoric nouns or pronouns, aliases and partial mentions to the specific entity they refer to.

For each entity a type, such as *person*, and a facet, such as *football player*, are also assigned and stored. This means that their search system supports retrieving quotations by, and also about, specific entities as well as certain categories, such as *doctors* or *politicians*. However, it is not possible to evaluate their quotation attribution extraction system since no scores are provided and only an execution time evaluation is mentioned. It is nonetheless possible to infer that it also suffers from very low recall, since they cite extracting sixty thousand quotations from fifty thousand news articles or blogs, i.e. little more than one quotation per text, while a much higher average of quotations per article should be expected in news.

In the ItAC (Sec. 2.1) corpus, an average of 9.22 attribution relations are found per article and even the partial annotation of attributions in the PDTB sum up to 4.5 per article. In addition, Bergler et al. (2004) report that up to 90% of the sentences in news language correspond to reported speech. The poor recall of the study by Liang et al. (2010) is not surprising, since attributions are retrieved only when signalled by one of a list of reporting verbs. In addition, precision could also suffer from the attempt to find a NE antecedent not only for each anaphoric pronoun but also for full noun mentions, while quotations can also be attributed to not-named entities, e.g. ‘scientists’, ‘a witness’.

In Alrahabi and Desclés (2008) and Alrahabi et al. (2010) the attribution of the quotation to its speaker is not addressed, but rather its identification and semantic categorization. The studies address only direct quotations that are accompanied by an expression of an act of speaking. They organize the quotation cues, which they call linguistic markers, in a semantic map of around sixty categories related to enunciative modalities. The studies then propose a rule-based system to identify quotations and annotate their semantic category in Arabic, French and subsequently also Korean. The system is based on approximately forty rules and 600-900 cues per language. It is tested on three more representative categories, taking only 15 quotation per category. The reported *precision* and *recall* are similar across languages and category and mostly around 80%-85%.

While most speaker attribution systems separately address coreference resolution and speaker attribution, Almeida et al. (2014) propose a joint model for direct quotations. The model treats quotations and mentions as nodes and builds a tree clustering together mentions referring to the same entity and quotations attributed to it. This is done by assigning a score to each arch linking two mentions or a mention to a quotation and scores depending on the paths in the tree. The features used are inspired by O’Keefe et al. (2012) and the system is trained and tested on a subset of the PARC 2.0 corpus developed as part of this thesis work (Sec. 3.3). The system identifies the speaker cluster of a quotation with 74% *F-score*.

Attribution is sometimes also included in coreference resolution studies, since direct quotations cause a shift in the pronouns used inside the quotation (e.g. *I* refers to the source and not the writer). In addition, entities mentioned inside the quotation, other than first person pronouns, usually cannot corefer with the source of that quotation. For example, Lee et al. (2011) uses a simple rule-based system to identify quotation speakers by retrieving the subject of reporting verbs in the proximity of a quotation. The identified speakers are then used to derive a set of heuristics (e.g. that the speaker and the mention *I* inside the quotation are coreferent).

2.2.1.2 Direct and Indirect Reported Speech

The studies presented in this section address the attribution of both direct and indirect reported speech. Two of them deal with languages other than English and have the goal of providing a way to structure and search information. The system presented in Sarmento and Nunes (2009), at the basis of the tool Verbatim (Sec. 2.3), works for the Portuguese language and is inspired by the NewsExplorer project (Pouliquen et al., 2007). The attribution extraction system is based on the retrieval of 35 possible reporting verbs and 19 patterns. It also extracts the quotation topic, but it does not handle pronoun and common noun source mentions. In the 570 quotations test set used, the system extracted incorrect NEs only in 1.8% of the cases. Similarly to the other studies relying on regular expressions and lists of verbs, it presents an extremely low recall. The authors estimate it to be around 5%. However, since they only retrieve one new quotation every 46 articles (about a third of the retrieved quotations, with the remainder identified as duplicates), their recall is likely even lower.

French news wires are the focus of the quotation extraction project SAPIENS[7] (de La Clergerie et al., 2009). This system verifies if the verb of the main clause is part of a list of 114 quotation verbs manually collected and selects its grammatical object as the attribution content and its subject as the source. Apart from attribution verbs, some prepositional phrases are also considered. The evaluation was performed on 40 manually sampled quotations. The relatively high *recall* of 80% they report, as opposed to the extremely low recall of previous works having a similar approach is not surprising, since the manually selected test set might be biased towards structures recognized by the system. On the retrieved attributions, they achieve a *precision* of 59% for content span and source identification. Most of the errors are due to a missing or incorrect source. Their novelty is the extension of the notion of quotation not only

[7] Within the SCRIBO project: `http://www.scribo.ws/`

to direct and indirect reported speech, but also to mixed reported speech, relatively frequently occurring in news language.

Attribution is also included in PICTOR (Schneider et al., 2010), a project for the development of a browser presenting quotations on a specific topic grouped by their subtopic and development over time. The quotation extraction system is based on a context-free grammar consisting of 273 non-terminal rules having numerical weights, manually assigned. Precision and recall of the extraction grammar have been measured both strictly, i.e. the match needs to be exact (56% *precision* and 52% *recall*) and by words correctly assigned to a quotation/speaker (86% *precision* and 75% *recall*).

Direct and indirect reported speech are also addressed by studies inspired by Bergler (1992), for the development of belief models to select information. They are based on the assumption (Bergler, 1992) that attribution is composed of a matrix clause, i.e. the attribution span, and a subordinate or complement clause, i.e. the content. The matrix clause contains at least the reporting verb, that is the only cue admitted, and the source, which has to be a subject NP. Based on these theoretical assumptions, Doandes (2003) developed an extraction system and tested it on a subset of the WSJ corpus, obtaining a *recall* of 44% and a *precision* of 92%.

In the same framework, Krestel et al. (2007) and Krestel et al. (2008) developed a quotation extraction system based on 6 general patterns, that they claim would match 97% of constructs using a pre-defined set of about 50 common reporting verbs. Tested on 7 articles from the WSJ containing 133 reported speech constructs, the system reports a *recall* of 79% and a *precision* of 100% for the detection of the reporting verb and source. Although their extraction system is not discussed in detail, from their error analysis it is possible to understand that it does not correctly handle circumstantial information because of the patterns limited flexibility. In Ex. (33), the circumstantial information in italics, which provides details about the context in which the quotation occurred, was incorrectly identified as part of the quotation.

(33) *Praising the economic penalties imposed by Congress last year*, he said it was "necessary to pursue the question of sanctions further". (Krestel et al., 2008, p.2827)

Moreover, the attribution structures it considers are limited by the theoretical assumptions described above, leaving out all attributions having no verb cue, verbs other than the 50 they list, source mentions not in the form of an NP or not in subject position (e.g. in case of a passive structure), implicit sources and contents that are not a grammatical sentence.

Ruppenhofer et al. (2010) presents a system for the identification of the speaker of the statements in German Cabinet Protocols. The rule-based system developed, automatically attributes a sentence to one or multiple speakers from a list of possible meeting participants. External source mentions account for about 14% of the speakers and are excluded. Since in German indirect reported speech is associated with subjunctive mood, this feature is exploited to identify sentences continuing reported speech from a previous sentence, hence inheriting its source. Otherwise, the source role is assigned to the subject of the main clause or the first NE of the sentence. The attribution system achieves 72% *precision* and 88% *recall*. Recall is considered more important for this task, since the attributions are retrieved to help historians identifying opinions by politicians. While all relevant statements should be returned by the system, erroneous attributions can be easily identified and discarded by the end user.

While most studies address the extraction of direct quotations only, Weiser and Watrin (2012) go in the opposite direction by defining an extraction methodology specifically for indirect quotations in French news texts. The study addresses only indirect quotations introduced by a speech verb. They make use of a grammar of 16 syntactic patterns and a list of reporting verbs. The system is evaluated relative to 2 patterns and 3 pre-selected verbs. The 140 spans the patterns identified were correct in around 74% of the cases.

Finally, Fernandes et al. (2011) presents a quotation extraction system for Portuguese, trained on a corpus of 802 quotations from GloboQuotes (Sec. 2.1.3). Their system has two components, separately addressing quotation extraction and quotation attribution. They model the quotation extraction task by training a system to identify the initial token of a quotation and then apply simple heuristics to identify the rest of the quotation. Quotations are then attributed to their source by training a model using PoS, quotation and coreference features. The overall system achieves 64% *precision* and 67% *recall*, while the speaker attribution component scores 79% *precision* and 79% *recall*.

2.2.2 Discourse Studies

Based on the representation of attribution in the RST Discourse Treebank (see Sec. 2.1.1), Skadhauge and Hardt (2005) developed a simple syntactic approach that can identify the kind of ARs annotated in the corpus. Using the available syntactic annotation from the PTB to identify attribution as the sentential complement of a verb, their system achieves 92% *F-score*. The high result is however limited to the type of ARs that are annotated in the RST corpus, which follow syntactic constraints.

Lin et al. (2014) developed an attribution span labeller as the final component of a discourse parser for the PDTB. The component identifies attribution spans, without associating them to a specific discourse relation previously identified. The labeller first splits the text into clauses by taking a syntactic approach, similar to the one proposed by Skadhauge and Hardt (2005). The following step is a binary classification of each clause into attribution span or non attribution span. The system achieves 79% partial and 66% strict *F-score*.

2.2.3 Opinion Studies

The second group of studies partly addressing the extraction of attribution relation originates in the framework of Opinion Analysis. Since the commercial demand in this field is particularly oriented towards determining the perception consumers have of a specific product or service, the identification of the opinion holder has long been neglected and seen as the least important feature of an opinion (Paroubek et al., 2010). While these studies deal with a great number of opinions, mainly from reviews, and express different point of views in terms of percentages, with no need to retrieve the specific source of each opinion, other applications of Opinion Analysis require the identification of specific sources. This is the case particularly for opinion-oriented summarisation and multi-perspective question answering tasks.

Since these studies focus on detecting opinionated or emotional language, their scope only partly overlaps with the attribution relations at focus in this thesis. While on the one hand their concern is also the attribution of opinionated language, on the other hand the attribution of statements is usually not included unless it is perceived as controversial. Moreover, these studies inlude the annotation of emotions, which are not in the scope of the current study. Nonetheless, because of the similarity of the task, it is of interest to provide an overview of approaches and achievements of these studies.

Opinion attribution extraction studies can be classified into three main groups, partially following the classification of approaches proposed by Xu et al. (2008), namely: approaches using FrameNet (Baker et al., 1998) or VerbNet (Schuler, 2005) which rely on semantic role labelling and heuristics; those based on dependency parsing and all other ones using Machine Learning classifiers with linguistic features.

2.2.3.1 Semantic-role Labelling Systems

The first group of approaches that tackle the identification of the source of opinions is based on semantic role labelling. These are inspired by Bethard et al. (2004) who semi-automatically annotated the opinion holders of opinion sentences extracted from FrameNet and PropBank (Kingsbury et al., 2002) exploiting their tendency to occupy the agentive role. Based on their data, only 10% of opinion holders were not agents.

Expanding on this approach, Kim and Hovy (2006a) collected frames from FrameNet for opinion verbs and adjectives and assigned semantic roles to the elements in the sentence. Finally, they selected the role corresponding to the opinion holder for each frame. While knowing the semantic relation between opinion holder and topic would be beneficial to attribution tasks, the results they obtained on a corpus of 100 sentences from news media text are rather low (47% *precision* and 34% *recall*), particularly because of the difficulty to set exact boundaries to the source mention. However, the scores show a significant improvement with respect to the baseline assigning the role of source to the subject of an opinion verb and of topic to its object.

A similar approach is developed by Das and Bandyopadhyay (2010) and employs emotion verbs from WordNet to extract sentences from VerbNet. The sentences are used to extract syntactic frames for each verb. Frames are then matched to the argument structure acquired through an independent rule-based system in order to assign the role of opinion holder. The results reported, although relatively high for this task, represent a rather small improvement with respect to the baseline, which classifies the verb subject as the opinion holder (*F-score* 67% vs. *F-score* 65%). These studies suggest that verb argument structure and semantic role labelling are not sufficient to boost opinion source extraction and other aspects of attribution need to be taken into account.

2.2.3.2 Dependency Parsing Systems

Since 2007 (Evans et al., 2007), the Multilingual Opinion Analysis Task at the NTCIR Workshop has introduced the sub-task of opinion holder identification. Participants have tested their systems on Chinese, Japanese and English news language. While results for the Chinese language are promising, very few participants have taken part in this task for the Japanese language and results for English are rather low, ranging from an F-score of about 2% to 44%. Scores are affected by the use of simplistic heuristics, like the identification of the source with the subject or agent of a communication verb (Bloom et al., 2007). However, attribution is expressed by a number of different structures and cues are not only verbs and especially not only reporting verbs.

After a first attempt with a heuristic rule based approach at NTCIR-7 (Seki et al., 2008), KLELAB proposed a Conditional Random Field (CRF) model (Lafferty et al., 2001) trained with the dependency tree connecting each word to the verb of a sentence and with the word probability to be in the phrase of the opinion holder. The F-score of the opinion holder extraction system for opinionated sentences was only around 37%.

Lu (2010) presented a system at NTCIR-7 based on dependency parsing, addressing the identification of opinion holders in Chinese news texts. The system makes use of reporting verbs as a feature for the identification of the opinion holder, which is assumed to be the subject. The opinion holder is then expanded in order to include attributional modifiers, quantifiers and other coordinate entities, e.g. in case of multiple sources. The system obtains around 68% exact match *F-score*.

2.2.3.3 ML-based Systems

The third group of studies includes the pioneer work by Choi et al. (2005) which uses Conditional Random Fields to identify the source of an opinion, emotion or sentiment expression based on the MPQA opinion annotation scheme. The study considers all possible expressions and not just verbs and achieves 54% *recall* and 72% *precision* for the exact match of the source. Also based on the MPQA corpus is the study by Kim and Hovy (2005) using Maximum Entropy to select the source from a list of all possible candidates identified. It considers syntactic features such as the syntactic path and the distance between each candidate and the expression to attribute. Subsequently, they apply their system to a corpus of German e-mails including pronouns as possible sources, the accuracy value they report dropping from 64 to 50% (Kim and Hovy, 2006b).

More recently, Wiegand and Klakow (2010) have developed an opinion holder extraction system using a sentiment lexicon from the MPQA and convolution kernels. The best results they report on a subset of the MPQA are 59% *precision* and 66% *recall* (94% accuracy). These scores were obtained by the combination of tree kernels based on constituency, sequence kernels having the span between the candidate source and the nearest predicate as scope and vector kernels using manually designed features.

Anaphora resolution was used by Kim et al. (2007) to address the identification of opinion holders in online news texts. Using a Support Vector Machine (SVM) classifier with a set of features, the system determines, for each opinionated sentence, if the source is *anaphoric*, *non-anaphoric* or the *author*. Different features are then used for each source group to develop a rule-based probabilistic model in order to select the actual opinion holder, i.e. the NE referent. Sentences, however, can contain more than one attribution and the low accuracy of this extraction system is due to being based on the assumption that sources are always NEs and that the anaphoric referent is expressed in the previous sentence, while it is in fact also found in the same sentence. While distinguishing between anaphoric and non-anaphoric sources may be an asset to the correct identification of the source referent, more precise rules should be derived to determine whether a common noun is to be considered anaphoric and a pronoun refers to a NE or a common noun.

2.2.4 Discussion

While none of the studies presented directly addresses the extraction of ARs to the extent proposed in this thesis, quotation and opinion studies present some similarities to the current task. Most quotation studies do not address the identification of quotations, but only their attribution to their speaker, which is similar to the identification of the source of a given AR content. While for direct ARs the content span can be relatively successfully identified with simple rules, this is not the case for the range of structures that can be used to express indirect and mixed ARs.

Lacking large-scale annotated resources and taking a limited scope approach to attribution, quotation extraction studies have developed rule-based approaches based on syntactic and semantic patterns that rely on lists of verbs. These suffer both from low recall and low precision and are not adequate for the identification of a wider range of ARs.

Opinion studies have developed systems to identify opinion expressions and link them to their opinion holder. These approaches do not identify the AR content but attribute opinion and sentiment expressions within it. The identification of opinion expressions is similar to the task of identifying AR cues and the identification of the opinion holder to source attribution. Some of these studies used the MPQA corpus to develop supervised systems.

Since quotation studies are closer to the current task, they will be used as a reference point to compare the approach in this thesis. While rule-based approaches represent a solid baseline for an extraction system, in Ch. 4 I propose a supervised model that can identify a broader range of attribution types and structures more reliably by learning from a large annotated corpus.

2.3 Applications

This section briefly presents applications available to date based on attribution. This overview will give an idea of the potential and relevance of attribution sensitive tools, particularly for opinion mining and information extraction tasks, but also show their current limitations, responsible for the still limited uses of these tools.

Quotes have been collected since remote times and the Internet has been populated by websites allowing to browse through famous quotations, particularly from writers and famous people, often grouped into topics (e.g. 'life', 'friendship', 'love'). One of these websites is ThinkExist[8], a collection of over 300,000 quotations submitted by thousands of individuals over several years. Since they are manually collected, these resources cannot be kept up to date and therefore are not suitable for retrieving recent information of the type normally found in news.

NewsExplorer[9] (Pouliquen et al., 2007) (see Section 2.2.3 for details) was developed with this exact purpose and is a rather popular tool (the authors report having over a million hits per day). It shows not only quotations by and about a NE source but also a basic profile and a list of entities related to it. The system works in several languages. The system recall is rather low and the precision is still not adequate, incomplete and incorrect attributions are in fact frequently found and this lowers the reliability of the tool. For example two of the six quotations listed in Fig. 2.1[10] are about and not by Osama bin Laden.

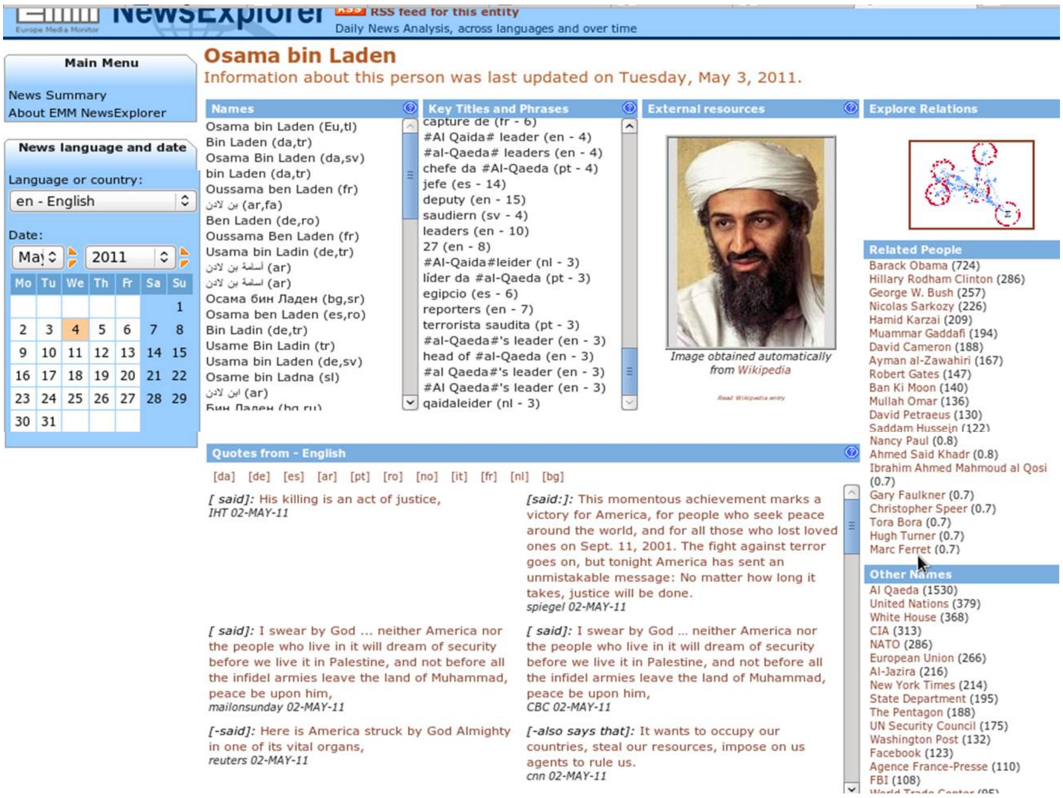
Google InQuotes[11] was an online tool allowing to browse through politicians' direct quotations about a set of specific topics and displayed two politicians on the same page in order to allow confrontations. The attribution extraction system was not described, however, it had a rather limited scope since only about 10-15 sources were supported for the US, Canada, India and the UK markets respectively. In addition, the topic of the quotation was simply based on the retrieval of the exact topic word inside

[8] `http://thinkexist.com/`

[9] `http://emm.newsexplorer.eu/`

[10] Screenshot taken from: `http://emm.newsexplorer.eu/`. Accessed on: 3rd of May 2015

[11] Launched in 2008 and formerly accessible at: `http://labs.google.com/inquotations/`. It was discontinued few years later.



Language or country: en - English

Date: May 2011

Mo	Tu	We	Th	Fr	Sa	Su
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

Osama bin Laden

Information about this person was last updated on Tuesday, May 3, 2011.

Names

Osama bin Laden (Eu,tl)
Bin Laden (da,tr)
Osama Bin Laden (da,sv)
bin Laden (da,tr)
Oussama ben Laden (fr)
بن لادن (ar,fa)
Ben Laden (de,ro)
Oussama Ben Laden (fr)
Usama bin Ladin (de,tr)
أسامة بن لادن (ar)
اسامة بن لادن (ar)
Осама бин Ладен (bg,sr)
Osama ben Laden (es,ro)
Bin Ladin (de,tr)
Usame Bin Ladin (tr)
Usama bin Laden (de,sv)
Osame bin Ladna (sl)
ابن لادن (ar)
Бин Ладен (bg,ru)

Key Titles and Phrases

capture de (fr - 6)
#Al Qaida# leader (en - 4)
#al-Qaeda# leaders (en - 4)
chefe da #Al-Qaeda (pt - 4)
jefe (es - 14)
deputy (en - 15)
saudiern (sv - 4)
leaders (en - 10)
27 (en - 8)
#Al-Qaida#leider (nl - 3)
líder da #al-Qaeda (pt - 3)
egipcio (es - 6)
reporters (en - 7)
terrorista saudita (pt - 3)
#al-Qaeda#'s leader (en - 3)
head of #al-Qaeda (en - 3)
#al Qaeda#'s leader (en - 3)
#Al Qaeda#'s leader (en - 3)
qaidaleider (nl - 3)

External resources

Image obtained automatically from Wikipedia

Explore Relations

Related People

Barack Obama (724)
Hillary Rodham Clinton (286)
George W. Bush (257)
Nicolas Sarkozy (226)
Hamid Karzai (209)
Muammar Gaddafi (194)
David Cameron (188)
Ayman al-Zawahiri (167)
Robert Gates (147)
Ban Ki Moon (140)
Mullah Omar (136)
David Petraeus (130)
Saddam Hussein (122)
Nancy Paul (0.8)
Ahmed Said Khadr (0.8)
Ibrahim Ahmed Mahmoud al Qosi (0.7)
Gary Faulkner (0.7)
Christopher Speer (0.7)
Tora Bora (0.7)
Hugh Turner (0.7)
Marc Ferret (0.7)

Other Names

Al Qaeda (1530)
United Nations (379)
White House (368)
CIA (313)
NATO (286)
European Union (266)
Al-Jazira (216)
New York Times (214)
State Department (195)
The Pentagon (188)
UN Security Council (175)
Washington Post (132)
Facebook (123)
Agence France-Presse (110)
FBI (108)
World Trade Center (95)

Quotes from - English

[da] [de] [es] [ar] [pt] [ro] [no] [it] [fr] [nl] [bg]

[said]: His killing is an act of justice,
IHT 02-MAY-11

[said]: I swear by God ... neither America nor the people who live in it will dream of security before we live it in Palestine, and not before all the infidel armies leave the land of Muhammad, peace be upon him,
mailonsunday 02-MAY-11

[-said]: Here is America struck by God Almighty in one of its vital organs,
reuters 02-MAY-11

[said:]: This momentous achievement marks a victory for America, for people who seek peace around the world, and for all those who lost loved ones on Sept. 11, 2001. The fight against terror goes on, but tonight America has sent an unmistakable message: No matter how long it takes, justice will be done.
spiegel 02-MAY-11

[said]: I swear by God ... neither America nor the people who live in it will dream of security before we live it in Palestine, and not before all the infidel armies leave the land of Muhammad, peace be upon him,
CBC 02-MAY-11

[-also says that]: It wants to occupy our countries, steal our resources, impose on us agents to rule us.
cnn 02-MAY-11

Figure 2.1: NewsExplorer screenshot.

the quote.

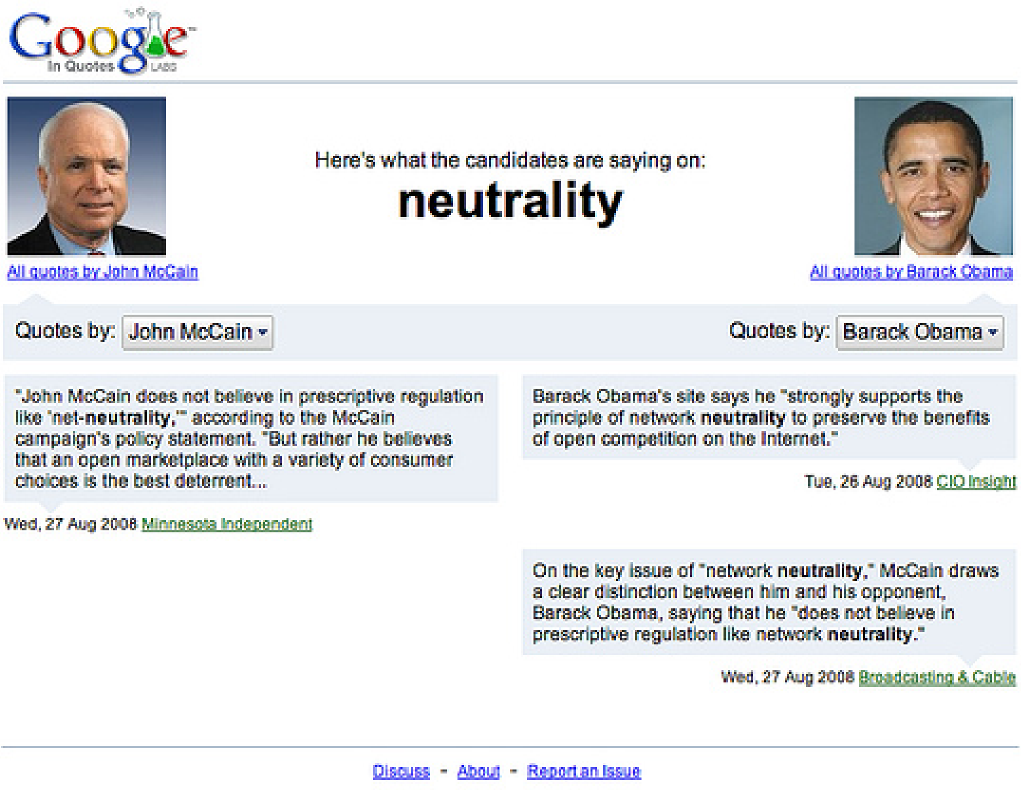
The example in Fig. 2.2[12] shows quotations containing the word ‘neutrality’, attributed to John McCain and Barack Obama. The second quotations on the right is however not correct. It is attributed to Barack Obama, who is the closest speaker to the attributive verb ‘saying’. However, the source would be ‘McCain’. Attribution structures can be rather complex and require more than a simple position based algorithm.

Verbatim is an application based on the system developed by (Sarmento and Nunes, 2009) (see Sec. 2.2.3). The initial system conflated into the SAPO VOXX web tool[13] (Fig. 2.3). The website displays quotations grouped by source or based on recency. It is possible to visualise the original online news provider or group of providers it was taken from. The system does not allow to search for a particular source or topic and suffers from an extremely low recall. SAPO VOXX is therefore not very reliable as it may miss relevant quotations and the information expressed in them.

The attribution extraction system proposed in Liang et al. (2010) was the basis for the creation of Evri, an online news search engine launched in 2008 and formerly

[12] Screenshot taken from `http://www.flickr.com/photos/rustybrick/2884309539/` Accessed on: 6th of March 2015.

[13] Accessible at: `http://voxx.sapo.pt/`


Google In Quotes LABS
All quotes by John McCain
All quotes by Barack Obama
Quotes by: John McCain
Quotes by: Barack Obama
Discuss - About - Report an Issue

Here's what the candidates are saying on:

neutrality

"John McCain does not believe in prescriptive regulation like 'net-**neutrality**,'" according to the McCain campaign's policy statement. "But rather he believes that an open marketplace with a variety of consumer choices is the best deterrent...

Wed, 27 Aug 2008 Minnesota Independent

Barack Obama's site says he "strongly supports the principle of network **neutrality** to preserve the benefits of open competition on the Internet."

Tue, 26 Aug 2008 CIO Insight

On the key issue of "network **neutrality**," McCain draws a clear distinction between him and his opponent, Barack Obama, saying that he "does not believe in prescriptive regulation like network **neutrality**."

Wed, 27 Aug 2008 Broadcasting & Cable

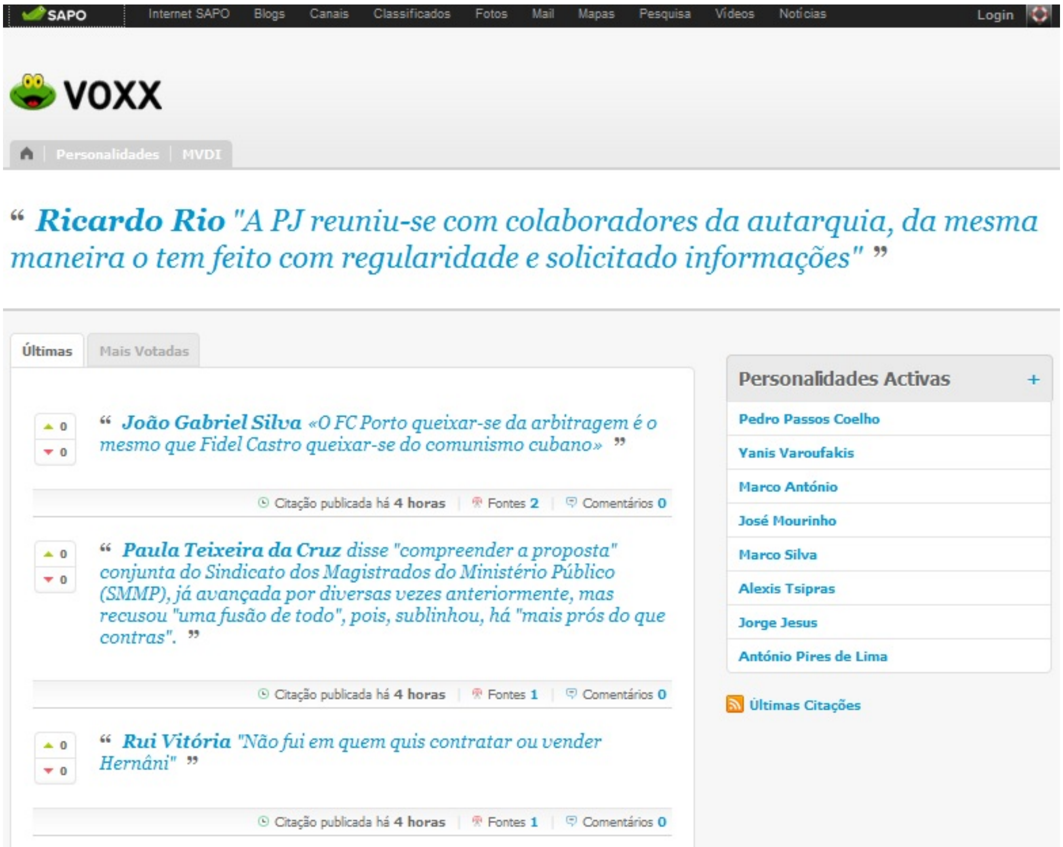
Figure 2.2: Google InQuotes screenshot.

available at `www.evri.com`. It was based on the idea to facilitate finding relevant breaking stories, one of the possible applications of attribution. Evri was then launched as a tablet application in 2011 and allowed to search for a specific topic and display relevant quotations and news, including where they were taken from. The company dissolved in 2012.

2.4 Conclusion

The available corpora annotated with some aspects of attribution are mostly small and incomplete and this hinders a deeper understanding of the structures carrying this relation and the development of supervised extraction systems. In order to have a suitable corpus to study attribution and develop extraction studies, I created a large and comprehensive resource, starting from the attribution annotations in the PDTB. Its development is described in Ch. 3.

The attribution extraction studies developed to date have addressed only a subset of attribution, identifying the most common structures of this relation. Most of the studies (for example Sarmento and Nunes (2009), Elson and McKeown (2010) and



“ **Ricardo Rio** "A PJ reuniu-se com colaboradores da autarquia, da mesma maneira o tem feito com regularidade e solicitado informações" ”

▲ 0 ▼ 0 “ **João Gabriel Silva** «O FC Porto queixar-se da arbitragem é o mesmo que Fidel Castro queixar-se do comunismo cubano» ”

Citação publicada há **4 horas** | Fontes **2** | Comentários **0**

▲ 0 ▼ 0 “ **Paula Teixeira da Cruz** disse "compreender a proposta" conjunta do Sindicato dos Magistrados do Ministério Público (SMMP), já avançada por diversas vezes anteriormente, mas recusou "uma fusão de todo", pois, sublinhou, há "mais prós do que contras". ”

Citação publicada há **4 horas** | Fontes **1** | Comentários **0**

▲ 0 ▼ 0 “ **Rui Vitória** "Não fui em quem quis contratar ou vender Hernâni" ”

Citação publicada há **4 horas** | Fontes **1** | Comentários **0**

Pedro Passos Coelho
Yanis Varoufakis
Marco António
José Mourinho
Marco Silva
Alexis Tsipras
Jorge Jesus
António Pires de Lima

Últimas Citações

Figure 2.3: SAPO VOXX screenshot.

Liang et al. (2010)) have based the recognition on the identification of the cue element in the text, thus showing its centrality and key function. While these studies have applied a number of techniques, from supervised classifiers (e.g. Choi et al. (2005), Kim and Hovy (2006b) and Wiegand and Klakow (2010)) to Semantic Role Labelling (Bethard et al. (2004) and Kim and Hovy (2006a)), the precision and recall of their systems remain unsatisfactory.

In order to boost attribution extraction recall, it is necessary to widen the spectrum of attribution structures addressed and reject incorrect assumptions, such as that cues are just verbs and can be identified with a list of common reporting verbs (Krestel et al. (2007); Sarmento and Nunes (2009)). Another misleading assumption is that the AR source always corresponds to an NE. This affects the possibility to correctly identify attribution to non-named entities. Finally, the annotation in the MPQA corpus has adopted an intra-sentential approach to attribution, which also affects the systems developed from it (e.g. Kim and Hovy (2005); Wiegand and Klakow (2010)).

None of the current available tools is sufficiently reliable to allow applications and users to make use of the extracted ARs knowing that these are correct and they will not miss precious information. Until we reach a satisfactory level of reliability of these tools, their applications will remain limited.

In Ch. 4 I will present a supervised AR extraction system that can identify source, cue and content of a wide range of ARs, including quotation and opinion attributions.

Chapter 3

A Corpus of Attribution Relations[1]

This chapter describes the project of creating a large corpus annotated with attribution relations, from the annotation scheme definition and validation to annotation and completion.

The annotation scheme, described in Sec. 3.1, was inspired by the annotation of ARs in the PDTB. The PDTB scheme was further extended and initially applied to ItAC, a pilot corpus of Italian (Pareti and Prodanof, 2010) described in Sec. 2.1.3.

Because the annotation of attribution was done by a single person, this aspect of the PDTB lacks of agreement scores. Since the annotation in this thesis aims at creating a reliable resource and is based on a modification of the PDTB annotation scheme, the first step was to prove the soundness of the proposed annotation. In order to validate the scheme, I have therefore conducted a preliminary inter-annotator agreement study, presented in Sec. 3.2.

In order to produce a complete and large resource for attribution, I first collected and extended the ARs annotated in the PDTB, adding specific labels for *source* and *cue* and separating the annotation from that of discourse connectives and their arguments. The set of ARs derived from the PDTB constitutes the two early versions of the corpus presented in Sec. 3.3. These two early versions of the corpus were used to conduct preliminary analysis of how ARs are encoded and to develop some of the attribution extraction models (see Ch. 4). However, these versions had a major drawback: only about half of the attributions were annotated, therefore the texts were a collection of labelled and unlabelled data.

A major manual annotation effort was undertaken in order to complete the corpus

[1] Parts of this chapter were published in Pareti and Prodanof (2010); Pareti (2012a) and Pareti (2012b).

as described in Sec. 3.4. With the help of three linguist annotators, the entire WSJ was fully annotated, over a period of four months, leading to the creation of PARC 3.0, a layer of annotation comprising almost 20k attribution relations. The corpus provides the data for the analysis of ARs reported in Sec. 3.5, which enables testing common assumptions in the literature.

3.1 Annotation Scheme

In order to produce a complete and large resource for attribution, I have further extended the annotation in the PDTB, adding specific labels for *source* and *cue* and separating the annotation from that of discourse connectives and their arguments. The annotation is based on a modification of the PDTB annotation scheme I previously developed for Italian (Pareti and Prodanof, 2010) and used to create the ItAC corpus (Sec. 2.1.3).

The annotation of ARs is lexically anchored and although different elements are annotated, all steps are performed at once for each AR and not in sequence. Once an AR is identified in the text, the annotators first mark its *cue*, i.e. an attributional verb or, less frequently, a preposition, a noun, an adverb or an adjective. The cue is then linked to the *source* element, unless implicit, and to the *content*, i.e. the text span perceived as attributed.

Optionally, information perceived as relevant for the interpretation of the attribution, completing or contributing to its meaning, can be marked and joined in the relation as *supplement*. This element was introduced to allow the inclusion of circumstantial information as well as additional sources (informers) (e.g. 'John knows FROM MARY ...)' or recipients (e.g. 'the restaurant manager told MS. LEVINE...' (wsj_1692).

Once an AR is annotated, feature values are selected. Six features were considered for inclusion into the scheme and tested through an inter-annotator agreement study. Four features correspond to those already proposed and included in the PDTB annotation (type, source type, determinacy and scopal polarity), another two are additional relevant aspects carried by the attribution, i.e. *authorial stance* and *source attitude*.

3.1.1 Elements

This section presents the elements that constitute an AR and how they have been included in the annotation. The chosen approach to attribution is lexicalized, i.e. it assumes that the elements are expressed in the text and the relation is anchored to textual expressions. Each AR element is therefore identified and annotated as a text span.

3.1.1.1 Cue

The cue is the element in the text that allows to establish the attribution relation and constitutes the link between source and content. The adopted approach assumes that for each AR there is one and only one cue. Therefore there has to be a textual element expressing the relation for the relation to exist and if two cues connect a source-content pair, they establish two ARs. This is the case in Ex. (34), where there are two ARs: a fact (A know B) and a belief (A believe B).

(34) **Analysts** know and believe *that the market is at a turning point.*

The cue is usually a verb, but nouns, prepositions or prepositional groups, possessives and adjectives and also adverbials can also have this function. Lexical cues can occur alongside punctuation clues, such as quotation marks and colon. In those cases, as in Ex. (35), where punctuation clues are the only cues in the text, they are annotated as the AR cue.

(35) **KIM**: *I got home, let the dogs into the house and noticed some sounds above my head, as if someone were walking on the roof, or upstairs. [...]* (wsj_1778)

While some verb-cues are expressed by reporting or opinion verbs semantically entailing the attribution relation, other verbs are not intrinsically attributional. This occurs predominantly with assertion attributions, since quotative constructions and punctuation clues allow for more flexibility on the verb choice. As discussed by Sams (2008), the quotative function of these verbs is activated by the construction they are in. She identifies two main relations holding between the quotative event and the verb: manner and co-temporal.

However, the distinction is not strict, since attributional cues lie on a continuum:

1. attributional verbs (e.g. ‘say’, ‘think’ and ‘know’)
2. attributional verbs entailing manner (e.g. ‘quip’, ‘grouse’, ‘tout’, ‘brag’ and ‘burble’)

3. manner verbs entailing a reporting verb (e.g. ‘beam’, ‘fume’ and ‘fret’)
4. verbs entailing the manner of an implicit co-temporal reporting verb (e.g. ‘sigh’, ‘shrug’ and ‘smile’) (Ex. (36))

Verbs that are merely co-temporal with the implicit attributional verb or occupy a verb-cue position but do not establish or imply the AR, as in Ex. (37), are not annotated as cues. In these cases we have to resort to punctuation to find a lexical anchor for the cue, even though also other elements, and in particular the attributional construction, contribute to establishing the relation.

(36) *Somewhere*, **the son** sighs, *things went terribly wrong with apartheid; today, whites even rely on blacks to police their separation.* (wsj_1760)

(37) Then **he** jumped into the market: *“I spent $30 million in the last half-hour.”* (wsj_2381)

Cues that are not intrinsically attributional are also those verbs that recall the previous cue by establishing a sequence in the narration such as ‘add’, ‘continue’ and ‘conclude’.

Cues are annotated together with their modifiers, since these can contribute to defining the relation. Adverbs in particular can contribute manner or authorial stance to the cue (e.g. ‘improperly ordered’, ‘vigorously oppose’ and ‘emphatically proclaims’). Negation particles may affect the factuality of the AR or reverse the polarity of its content.

3.1.1.2 Source

The **source** is annotated as the text span where the source is mentioned. Grammatically, sources are usually expressed by a proper noun, a common noun or a pronoun and annotated together with the rest of the noun phrase they are part of, thus including modifiers, appositives and relative clauses. Rarely, when the cue is a noun, the source can be expressed by an adjective (e.g. ‘the **presidential** statement’).

Semantically, sources can be named as well as not named entities such as ‘a witness’ and ‘the company’, specific or generic such as ‘analysts’ and ‘most people’, pluralities and metonymic referents such as ‘Washington’, ‘the White House’, ‘the office’, ‘the letter’. Sources might also be implicit and are therefore an optional element in the annotation. Implicit sources are not only associated with passive attributional structures, but also impersonal constructions with the cue verb in the infinitive (Ex. (38)) or gerund form.

(38) “Just to say *the distribution system is wrong* doesn’t mean anything,” [...] (wsj_0082)

3.1.1.3 Content

The *content* is the span of text corresponding to what is attributed. In principle any span of text can be the content of an AR. Commonly, this will be a clause. However, it can also be a single word, a phrase, one or more sentences or paragraphs. This happens in particular with direct ARs having quotation marks as delimiters, since they allow for more flexibility with contents stretching also intersententially.

The content span might also be discontinuous, since source and cue can appear interpolated within it as a parenthetical construction as in Ex. (39), or the content span can resume in a contiguous sentence without any further clues being required as in Ex. (40).

(39) *Today*, **he** frets, *exports and business investment spending may be insufficient to pick up the slack if stock prices sink this week and if consumers retrench in reaction.* (wsj_2397)

(40) *“The Caterpillar people aren’t too happy when they see their equipment used like that,”* shrugs **Mr. George**. *“They figure it’s not a very good advert.”* (wsj_1121)

Unlike the source, the content element cannot be implicit. However, it can be expressed by an anaphoric pronoun (e.g. the cataphoric content in Ex. (41)). In other cases, the content is not present but simply alluded (e.g. *He said the truth/ two words/ what he had to say*). Those apparent ARs are not annotated since the text span corresponding to the content is not present, not even anaphorically.

(41) Although **Paribas** denies *it*, analysts say the new bid in part simply reflects the continuing rivalry between France’s two largest investment banking groups. (wsj_1319)

3.1.1.4 Supplement

Beside the constitutive elements of ARs, the surrounding context can carry further information relevant to the AR. When the attribution span contains relevant elements that are neither part of the source nor of the cue, these should be marked as SUPPLEMENT. In particular, supplemental elements are those providing a context for interpreting an AR including its:

- Setting (time, place, audience) as in Ex. (42)[2].
- Topic as in Ex. (43).
- Communication medium as in Ex. (44).
- Relevance to the author's argument as in Ex. (45).
- Manner as in Ex. (46).

(42) *"Ideas are going over borders, and there's no SDI ideological weapon that can shoot them down,"* **he** told [A GROUP OF AMERICANS] [AT THE U.S. EMBASSY] [ON WEDNESDAY]. (wsj_0093)

(43) OF SONY, **Mr. Kaye** says: *"They know there's no way for them to lose. They just keep digging me in deeper until I reach the point where I give up and go away."* (wsj_2418)

(44) **Trade and Supply Minister Gerhard Briksa** said IN A LETTER PUBLISHED IN THE YOUTH DAILY JUNGE WELT *that the rise in alcohol consumption in East Germany had been halted*; (wsj_1467)

(45) AS AN INDICATOR OF THE TIGHT GRAIN SUPPLY SITUATION IN THE U.S., **market analysts** said *that late Tuesday the Chinese government, which often buys U.S. grains in quantity, turned instead to Britain to buy 500,000 metric tons of wheat.* (wsj_0155)

(46) *"A very striking illusion,"* **Mr. Hyman** says [NOW], [HIS VOICE DRIPPING WITH SKEPTICISM], *"but an illusion nevertheless."* (wsj_0413)

The information contained in the supplement might still not be sufficient to fully evaluate and fully understand an AR. In Ex. (46) we do not know what the source considers an 'illusion', i.e. the topic this assertion is about. Nonetheless, the supplement usually provides sufficient elements for the interpretation of the AR.

[2] Supplements are represented in the examples in small capitals.

3.1.2 Features

tures can be worth including in the annotation of ARs. These are the features relative to aspects that contribute to the interpretation of the information conveyed by the content and to determine the inherent reliability of the AR. Eight main features of attribution have been identified and included in the annotation scheme. Four of them are derived from the ones included in the PDTB annotation and described in Sec. 2.1.1. These are: attribution type (Sec. 3.1.2.1); source type (Sec. 3.1.2.2); factuality and scopal change (Sec. 3.1.2.3).

Two other features have been introduced to account for ARs in which the authorial stance (Sec. 3.1.2.4) and the attitude the source expresses towards the content (Sec. 3.1.2.5) are also expressed. While the most frequent reporting verbs, such as 'say', tend to be more neutral and therefore less informative, less frequent verb-cues, particularly those not normally associated with a reporting meaning, often provide additional information. For example, manner verbs such as smile, chuckle, purr and sniff can express the source's attitude towards the content. The authorial stance, namely the author's commitment towards the truth of the content, is also mostly expressed by the choice of verb: committed, such as 'acknowledge' and 'admit'; not expressing any commitment, such as 'say' and 'announce'; not committed, such as 'lie' and 'joke'.

For both features, while it is possible to pre-classify some verbs to help the annotator, it is not possible to have an exhaustive list (otherwise this features could be automatically derived). A list could only provide an inventory of more prototypical cases, leaving out most of the challenging borderline ones. In addition, it would assume the context to be irrelevant, while it also concurs to determine the feature value (e.g. 'say' reflects a neutral attitude while 'say with a smile' or 'doubters say' are not neutral).

Finally, there are two features that do not require annotation as they can be reliably computed: the quote status and the level of nesting. The quote status (Sec. 3.1.2.6) accounts for the attributed material being directly quoted, partly directly quoted or indirectly reported. Only assertion ARs or quotations can be direct or partly direct. The level of nesting (Sec. 3.1.2.7) is a measure of how many ARs contain a given AR. This affects the reliability of the AR and of the information conveyed by its content, which potentially underwent more manipulation.

3.1.2.1 Attribution Type

The PDTB annotates the **attribution type**, according to the taxonomy of abstract objects described by Asher (1993). This distinguishes between word immanent objects, represented by *events* and *states*, and purely abstract objects, i.e. *propositions*. The two categories are the extremes of a continuum of world immanence, while *facts* occupy an intermediate position, having some traits in common with both events and states. In the PDTB annotation, propositions are further divided into *assertions* and *beliefs*. The attribution type reflects the commitment of the source towards the abstract object expressed by the content span.

Verbs of communications are derived from the groupings proposed by Levin (1993), with assertions corresponding to 'assertive predicates or verbs of communication' such as 'announce', 'observe', 'reveal', 'suggest' and 'claim'. Although very useful in principle, a classification irrespective of the natural context of occurrence of each verb has big limitations. While 'suggest' indeed expresses an assertion in Ex. (47), it conveys an eventuality in Ex. (48), having the intent of influencing the hearer. Similarly, 'observe' could also express a fact.

(47) **Economists** suggested *that if the pound falls much below 2.90 marks, the government will be forced to increase rates to 16%, . . .* (wsj_1500)

(48) **Mr. Canelo** suggests *that investors compare price/earnings ratios* (the price of a share of stock divided by a company's per-share earnings for a 12-month period) *with projected growth rates.* (wsj_1761)

The taxonomy of eventualities is derived from Sag and Pollard (1991), where non–exhaustive lists of verbs of commitment, influence or orientation are provided. Also in this case, however, only the context can tell if, for example, 'agree' is an eventuality (e.g. agree to do something) or an expression of opinion (e.g. agree with someone's belief). The distinction between eventualities and beliefs remains subtle. A sentence like 'I believe it won't rain tomorrow'could be perceived as expressing both a belief or an expectation.

3.1.2.2 Source Type

Another feature annotated in the PDTB accounts for the **source type**: *writer*, *other* or *arbitrary*. Defining whether the source is specific (other) or generic (arbitrary) is relevant: In most cases it is possible to disambiguate and resolve the source to an actual entity for type other, in case of arbitrary sources this is not viable. The definiteness of the source is a continuum and contextual information is required. For example, a source like ‘everyone’ can express a generally accepted view (arbitrary) or the shared view of each member of a specific group (other) (e.g. ‘everyone in the House of Lords’ meaning the ministers). As noted in previous work (Pareti and Prodanof, 2010), in case of multiple sources, these might have conflicting types, as in ‘My assessment and everyone’s assessment is ... ’ (wsj_2012). For those cases, a new value for the source type was added to the scheme: ‘mixed’.

3.1.2.3 Factuality and Scopal Change

The last two features annotated in the PDTB are *determinacy* and *scopal polarity*. Determinacy accounts for the factuality of the attribution itself, i.e. if the relation between source and content is presented as a fact of the real world or an unreal or hypothetical fact. Scopal polarity on the other hand marks if a negation, apparently scoping over an attributional verb (e.g. ‘didn’t say’, ‘deny’) instead reverses the polarity of the attributed content.

Following the terminology introduced in Pareti and Prodanof (2010), determinacy is renamed to **factuality** and scopal polarity is redefined as a change in scope not exclusively bound to a polarity shift and referred to as **scopal change**. This was introduced in order to include more than one element under the same feature all of which affect the factuality (other than negation particles) that could shift their scope from the relation to its content (e.g. ‘if’).

Although in theory determinacy and scopal polarity are complementary, real language can be ambiguous with respect to the attribution being presented as non–factual or as having a content with inverted polarity. An AR such as the one in Ex. (49) could be interpreted either as non-factual or as an expectation that ‘the merger will not face any regulatory hurdles’ (scopal change).

(49) **They** don’t expect *the merger to face any regulatory hurdles.* (wsj_1660)

Apart from modal verbs or particles directly affecting the verb-cue, the factuality can be determined by the verb-cue mode (e.g. conditional, imperative) and tense (e.g. future). However, the AR factuality is not exclusively expressed on the cue. The source (Ex. (50)) as well as an interrogative structure can also make the AR non factual.

(50) **No one in his right mind** actually believes *that we all have an equal academic potential.* (wsj_1286)

3.1.2.4 Authorial Stance

Authorial stance is a relevant feature carried by attribution that is worth including in the annotation. Unlike the attribution type, which reflects the source's commitment, the authorial stance reflects the author's commitment towards the truth of the AR content, and is considered to be the expression of the reporter's voice (Murphy, 2005) and their beliefs (Diab et al., 2009). As noted by Kessler (2008), mentioning does not imply agreeing. On the contrary, Thompson and Yiyun (1991) observe that the choice of non reporting (i.e. not attributing the content to a different source) implies a positive evaluation, since the author takes direct responsibility and commits to its truth.

The concept of author is a relative concept. The author of an AR is in principle the above source (i.e. the author of the text or another source). While this can be assumed for assertion ARs, for other types of nested ARs it is not as clear. In Ex. (51), the commitment towards the truth value of the AR content 'X' is expressed by the choice of 'admit' as AR cue. This choice could be an addition of the author of the text and not correspond to what 'John', the source of the including AR, wanted to express.

(51) a. John said that Mary admits X. (John is the author suggesting that X is factual)

b. John wants Mary to admit X. (X being factual could be expressing John's as well as the writer's belief)

The annotation distinguishes between neutral (e.g. 'say'), committed (e.g. 'admit') or non–committed (e.g. 'lie' and 'joke') authorial stance. While the authorial stance can be expressed by the verb choice, it can also be expressed by other elements. Beside the choice of cue, Kessler (2008) identifies also the source. Choosing to mention the source with a negative term such as 'nobody', 'fools' and 'idiots' the author conveys also her disagreement with what is conveyed by the content.

3.1.2.5 Source Attitude

The **source attitude** reflects whether a sentiment is associated with the attitude the source holds towards the content. The annotation scheme allows for five different values: positive (e.g. 'beam', 'support', 'encourage', 'hail' and 'brag'), negative (e.g. 'shout', 'decry', 'fume' and 'convict'), tentative (e.g. 'believe', 'ponder' and 'sense'), neutral (e.g. 'report') or other (in case a sentiment is expressed which does not fall into any of the previous categories).

Similar to the authorial stance, the source attitude is determined in context. Some cue verbs might have a clear sentiment associated, but most are relatively neutral and the sentiment may be determined by additional elements. The semantic content of the AR may be misleading for the annotators since they could consider an AR as positive or negative based on whether what the content expresses is positive or negative.

3.1.2.6 Quote Status

The **quote status** feature identifies whether an AR is a *direct* or partly direct (called *mixed*) quotation or it is *indirectly* reported. Since only speech acts can be directly reported, as they allow for verbatim of the original words uttered, only ARs of type assertion can be direct, indirect or mixed. All other types (beliefs, facts and eventualities) are necessarily indirect. Few borderline cases exist, such as Ex. (52), where the content is attributed as an eventuality, however, part of it is verbatim, or Ex. (53), where the verb-cue suggests that a belief is attributed, however, this is expressed by a direct quotation.

(52) Similarly, **Rick Wamre, a 31–year–old asset manager for a Dallas real–estate firm**, <u>would like</u> *to see program trading disappear because "I can't see that it does anything for the market or the country."* (wsj_0121)

(53) **Takuma Yamamoto, president of Fujitsu Ltd.**, <u>believes</u> *"the 'money worship' among young people . . . caused the problem."* (wsj_0094)

For the annotation, the quote status should be determined based on the content only and not on the choice of the reporting verb or the semantic of the content. The quote status of an AR is assigned by an algorithm (see Sec. 4.1.3) that considers quotation marks. If the complete content span lies within quotation marks it is assigned the value 'direct', if there are quotation marks but not at the edges of the span it is labelled as 'mixed' and if there are no quotation marks in the span it is recognized as 'indirect'.

3.1.2.7 Level of Nesting

One of the characteristics of attribution is that an AR can also occur inside another AR. When this happens, the AR is **nested**. The idea of applying the concept of nesting to attribution is inspired by the annotation of opinions and emotions presented in Wiebe et al. (2005). In their work, nesting is annotated on the source, by listing all sources, including the writer, as in Ex. (54), which they provide with respect to the expression 'criticism'.

(54) Source: writer, Foreign Ministry, U.S. State Department

Text including the sources: The foreign ministry said it was surprised, to put it mildly, by the U.S. State Department's criticism ... (Wiebe et al., 2005, p.14)

Listing all sources is not necessary for ARs, since their content span is part of the annotation and nesting can therefore be analysed as the inclusion of an AR into the content of another, as Ex. (55) shows.

(55) $_{1st}$[*Moreover*], **Mr. Guber** claims, $_{1st}$[***Mr. Semel*** *told him* $_{2nd}$*[that Mr. Ross probably wouldn't object "if it were anybody other than Sony. But Sony is a problem."]]* (wsj_0578)

A nested AR inherits from the embedding one not only the source, but also its relation with the content, i.e. the attitude it holds towards it. In Ex. (56), the content of the nested AR 'she will come back' is affected by both sources (Mary and John) and their trustworthiness. However, in Ex. (56a), the writer presents the attitude of the first-level source as uncertain and a belief, while in Ex. (56b) she presents it as factual and as constituting an assertion.

(56) a. John doubts that Mary said she will come back.

b. John announced that Mary said she will come back.

The level of nesting of an AR can be computed by taking its cue span and verifying if it is part of the content span of another AR (Algorithm 1). For each AR content the cue span is part of, the level of nesting is increased by one. A level of nesting of one corresponds to ARs directly inserted into the text with the relation created by the author of the text. These ARs will be referred to as *first-level* or not nested. A level of nesting of two or more corresponds to ARs that have been explicitly made by another entity and are inserted into the content span of one or more other ARs. These attributions are referred to as *nested*.

Algorithm 1 Compute Level of Nesting of an AR

```
procedure GETLEVELOFNESTING($AR_1$)
    $AR_1$ level of nesting = 1
    for AR in document do
        if $AR_1$ cue span in AR content span then
            $AR_1$ level of nesting + 1
```

3.2 Validating the Schema

The approach to the annotation of ARs presented in this thesis (Sec. 3.1) is inspired by the PDTB scheme. This section describes an inter-annotator agreement study that was conducted in order to verify the validity of the PDTB derived annotations before employing this resource for the development and testing of attribution extraction studies. The study also evaluates the applicability of the proposed annotation scheme before applying it to further annotate the corpus.

3.2.1 Study Definition

In order to test the annotation scheme and identify problematic aspects, a preliminary inter–annotator agreement study was developed on a sample of the WSJ corpus. This sub–corpus consists of 14 articles, selected in order to present instances of all possible attribution types and feature values. Two expert annotators were independently asked to annotate the articles using the MMAX2 annotation tool (Müller and Strube, 2006), following the instructions provided in an annotation manual (see Appendix A).

The guidelines make use of surface clues to guide the annotation. Attribution is lexicalized as having a textual anchor, the *cue*, that represents the starting point of the annotation. *Source*, *content* and *supplement* should be subsequently identified and the relative span annotated according to given rules. For example, the source span should represent the full source mention including all modifiers, e.g. appositives and relative clauses, but anaphoric sources should be annotated without resolving the anaphora. For most other aspects of the annotation, and in particular for the annotation of features, conceptual instructions are provided together with a list of prototypical as well as borderline examples. Since the feature values lie on a continuum between the allowed distinctions, the annotators are invited to make a decision in context, according to their interpretation of the underlying principles.

Preliminary training was conducted to familiarise the annotators with the tool and with the annotation scheme. During the training phase, the annotators independently annotated one article, and then confronted their result and were able to discuss problematic ARs. At this stage, additional guidance concerning uncertain cases and systematic errors was provided.

3.2.2 Results

3.2.2.1 Attribution Relation Identification

The annotators identified 380 attributions in common of the overall 491 ARs they annotated. This corresponds to an average of 35 ARs per article. Since they were annotating different text spans, the agreement was calculated using the *agr* metric proposed in Wiebe et al. (2005). The *agr* metric is a directed agreement score that can be applied to relation identification tasks where the annotators do not choose between labels for a given annotation unit, but have to decide whether there is a relation and if so, the scope of the text span that is part of it. For two given annotators a and b and the respective set of annotations A and B the annotators performed, the score returns the proportion of annotations A that were also identified by annotator b.

$$agr(a||b) = \frac{|A \cap B|}{|A|} \tag{3.1}$$

$$agr(b||a) = \frac{|A \cap B|}{|B|} \tag{3.2}$$

$$agr_{ab} = \frac{agr(a||b) + agr(b||a)}{2} \tag{3.3}$$

For the AR identification task, the *agr* metric was 0.87. This value reflects the proportion of commonly annotated relations with respect to the overall relations identified by annotator a and annotator b respectively (i.e. the arithmetic mean of $agr(a||b)$ 0.94 and $agr(b||a)$ 0.80).

The disagreement for this task was mainly caused by the tendency of one annotator to consider some expressions of sentiment as attribution, although these are not in the scope of this project, as well as several cases where there is no explicit attribution even though the future tense can be perceived as expressing an intention as in Ex. (57). These errors can partly be corrected with additional training.

(57) Yet CBS will air only 12 regular-season games, 26 fewer than ABC and NBC. (wsj_1057)

Higher disagreement can help identifying what are less prototypical ARs. Less disagreement occurs when all three constitutive elements are explicitly expressed and take a more common structure. A number of other structures that occur less frequently are instead more challenging to identify. In particular, more than a third of the disagreement occurred with ARs having a content span expressed by a noun phrase as in Ex. (58). Another problematic distinction in some cases was that between attributions expressing a sentiment or an opinion or simple expressions of sentiment or opinion. In Ex. (59) the identified AR is just an expression of opinion and there is not a real content span, this is just the target the opinion is about. In these cases the content is usually identified with a prepositional phrase. Other sources of disagreement occurred with ARs having an implicit source or certain verbs, such as 'call' and 'name' as in Ex. (60) which might entail a speech act, albeit recurrent.

(58) **Wilder** has managed to get across *the idea that Coleman will say anything to get elected governor*. (wsj_0041)

(59) [...] **he** had some concerns *about the language in the legislation* (wsj_0041)

(60) Despite all these innovations, most of the diamonds are still found in the sand swept away by the men wielding shovels and brushes – the ignominiously named *"bedrock sweepers"* who toil in the wake of the excavators. (wsj_1121)

Higher disagreement correlates with the identification of nested attributions. This can be in part attributed to the characteristic of the annotation tool: once an attribution is annotated, it is more difficult to visualise another attribution expressed in its content span, as it is already marked and therefore less visible. Moreover, nested attributions are shorter, do not rely on punctuation clues as they are rarely direct and have a higher proportion of types other than assertions. While overall 22% of the ARs identified by the annotators are nested, the proportion drops to 15.5% for the ARs identified by both annotators. Nested ARs represent instead over 44% of the ARs identified only by one annotator.

3.2.2.2 Span Selection

The agreement with respect to choosing the same boundaries for the text span to annotate was also evaluated with the *agr* metric. The results (Table 3.1) are very satisfactory concerning the selection of the spans for the source (.94 *agr*), cue (.97 *agr*) and content (.95 *agr*) elements. Since supplemental information was only annotated for less

Cue	Source	Content	Supplement
0.97	0.94	0.95	0.37

Table 3.1: Span selection *agr* metrics.

than 1 in 4 attributions, the *agr* was calculated only for the relations where at least one annotator identified a supplement. The low agreement of .37 shows that what constitutes material to complete the attribution or is relevant to its understanding is rather subjective. The annotation of supplemental information was included in the study as exploratory, in order to give the annotators a label for what they considered part of the AR or relevant but would not represent one of its constitutive elements.

3.2.2.3 Features Selection

Once an AR was identified, the annotators were asked to select the values for each of the six annotated features. Several issues emerged from this task. Despite very high percentage agreement values (see Table 3.2), the corrected Kappa measure shows a different picture with results in part not satisfactory. The selection of the *source type* and the *factuality* value are above the 0.67 recognised by some literature as the threshold allowing for some tentative conclusions, as discussed in detail by Artstein and Poesio (2008). *Type* and *scopal change* are also above 0.6.

On the other hand, the two newly introduced features of *authorial stance* and *source attitude* reached only .20 and .48 Kappa agreement respectively. Even the attribution type had relatively low agreement (.64 Kappa). On the contrary, the percentage agreement is very high due to the fact that the values the feature can have are extremely imbalanced, with certain values being predominant and others rare. Table 3.2 reports percentage and Kappa agreements, as well as the number of instances the annotators disagreed on out of the 380 commonly annotated ARs.

3.2.3 Disagreement

This section will present the analysis of the disagreement concerning the feature selection and analyse if this is attributable to the annotation scheme, the data or rather to the way the annotation was performed.

The attribution type appeared to be a very problematic feature, since attributional verbs can belong to more than one category depending on the context, but also on the

Features	Percentage Agreement	Cohen's Kappa	N Disagreements
Type	0.83	0.64	63
Source	0.95	0.71	19
Scopal change	0.98	0.61	5
Authorial stance	0.94	0.20	21
Source attitude	0.82	0.48	67
Factuality	0.97	0.73	9

Table 3.2: Percentage and Kappa agreement values for the selection of AR features. The final column reports the absolute number of disagreements for that feature out of 380 commonly identified ARs.

way this is interpreted. The confusion matrix in Table 3.3 shows that most of the uncertainty involved eventualities and facts. Not only verbs like 'see', having different readings depending on the context (i.e. see can be used to express a perception (factual) as well as an opinion (belief)), led to disagreement. Several verbs appear to be intrinsically ambiguous. In Ex. (61) 'expect'was perceived by one annotator as entailing a belief and by the other annotator as entailing an attempt to influence the hearer and thus as an eventuality.

AR Type	Assertion	Eventuality	Fact	Belief	Tot.
Assertion	**248**	30	4	9	291
Eventuality	2	**28**	1	3	34
Fact	3	1	**7**	5	16
Belief	1	4	0	**34**	39
Tot.	254	63	12	51	380

Table 3.3: Confusion Matrix for the annotation of the AR type feature.

(61) However, in interviews later, both ministers stressed that **they** expect *future OPEC quotas to be based mainly on the production capacity and reserves of each member*. (wsj_1428)

Eventualities were mostly confused with assertions. In Table 3.3 we can see that one annotator identified almost twice as many eventualities with respect to the other annotator, 63 vs. 34. These were verbs such as 'agree', 'suggest', 'insist' and 'warn' which can be seen as verbs of communication, thus assertions, as well as commitment or influence, thus eventualities, depending on the context and the subjective interpretation of the annotator.

Determining the source type caused lower disagreement. Some causes of errors originated from less intuitive constructions such as passive forms, where the source is usually not explicitly expressed, and the annotator is required to judge the type of the implicit referent. Source type was also ambiguous in interviews, where the interviewee made attributions to a non better specified *you* which could be intended as a reference to the interviewer (supposedly the writer) as well as another non specified entity that was present during the interview or even an impersonal *you*.

In addition, real sources lie on a continuum between specific referents (named entities) and generic entities (e.g. 'people' and 'one') 'indicated via a non specific reference '(Prasad et al., 2006, p. 33). While 'some OPEC sources' is more specific than 'rumours', is it specific enough to be classified as *other*? One way to drive the annotation would be to provide a test to assess if the source has a specific referent in the real world.

The factuality of the attribution presented less complexity. Ambiguity arose, however, with conditionals as in Ex. (62) and past tenses. One annotator interpreted the attribution of intentions in the past as implying that they are no longer factual. The annotators were also unsure whether to annotate an attribution in the scope of a negation as non–factual or rather presenting a scopal change, in particular for belief ARs such as Ex. (63). The issue here is that a negated attribution can be indeed factual while implying the negation of its content.

(62) Mr. Nazer, the Saudi oil minister, reiterated here that **the kingdom** would insist *on maintaining its percentage share of OPEC production under any quota revisions*. (wsj_1428)

(63) "**I** don't think *I have a life style that is, frankly, so flamboyant*," he says. (wsj_2113)

Probably the main reason causing the source attitude and authorial stance features to 'fail' the agreement test is that values for these features are extremely imbalanced (see Table 3.4). The vast majority of ARs have neutral values for these features (365/380 for the stance and 275/380 for the attitude). The infrequency of values other than neutral makes it challenging for the annotator to identify these cases while already confronted with a complex annotation task and to reach internal consistency (e.g. 'introduced' was perceived in a similar context both as suggesting *neutral* and *committed* stance by the same annotator). Moreover, the monotony of always having the same values could lower the annotators' attention and make them more prone to forgetting to change the default values.

AR Source Attitude	Neutral	Positive	Critical	Other	Tentative	Tot.
Neutral	**275**	7	10	1	25	318
Positive	8	**10**	2	1	1	22
Critical	5	0	**28**	0	0	33
Other	4	0	0	**0**	1	5
Tentative	1	0	1	0	**0**	2
Tot.	293	17	41	2	27	380

Table 3.4: Confusion Matrix for the annotation of the AR source attitude feature.

3.2.4 Discussion

This section reports on the challenges that arose from the pilot annotation using the proposed scheme for attribution relations. The agreement study showed that attribution is a relatively well defined relation and that there is little disagreement on determining the span corresponding to its constitutive elements *cue*, *source* and *content*.

The results highlighted some unsolved problems concerning the proposed features. In particular, the need for a better identification of the boundaries for values on a continuum such as the attribution type, and the potential overlap of the *determinacy* and *scopal polarity* features. On the one hand, the scale of disagreement might reflect the complexity of the particular task, on the other, further experiments would be needed to exclude or reduce the effects deriving from the particular set up of the annotation task and the data distribution.

One difficulty in applying the proposed annotation schema originated from the number of elements and features that needed to be considered for the annotation of each attribution. This suggests that by decreasing its complexity, the number of errors could be reduced. The annotation should be therefore split into two separate task: the AR annotation and the feature selection. This way the annotators would be faced with less decisions at a time.

For decisions such as the attribution factuality and whether the scope of the negation affects the content instead of the AR itself, test questions could be a useful strategy to ensure a better convergence of the results.

While a redefinition of some of the feature is desirable in order to reduce ambiguity and subjectivity, the low agreement is greatly affected by the imbalanced data. In order to test the features complexity one possibility would be to select a balanced subset of the corpus that contains a similar number of instances for each feature value. However, this assumes the values to be known beforehand and it would not be representative of the corpus distribution where some feature values are indeed predominant and some rarely occurring.

It is highly desirable to build a complete resource for attribution studies enriched by relevant features that affect the interpretation and perception of ARs. However, in the light of the inter-annotator agreement study, I decided to restrict further annotation efforts to the AR span selection and leave the annotation of the features to future work.

3.3 Early Versions of the Corpus

3.3.1 Data Collection

The attribution corpus described in this section was created starting from collecting the attributions annotated in the PDTB. In this resource, each discourse connective and its two arguments are associated with an attribution span, i.e. the span of text where the attribution relation is established. The annotation comprises also some features as presented in Sec. 2.1.1.

Since the content of a newspaper article is attributed to its writer by default, unless otherwise expressed, such ARs have been excluded from the collected data. Each AR had to be reconstructed by joining one or more discourse connectives and arguments having the same attribution span into a same content span. The example in Fig. 3.1 illustrates the PDTB annotation of two discourse connective and relative arguments corresponding to the attribution relation in Ex. (64). The attribution span is reported in the second *Text* field of the discourse connective, while the content of the attribution is fragmented, as it comprises the argument texts of both discourse connectives and the explicit discourse connective itself.

(64) *"There's no question that some of those workers and managers contracted asbestos–related diseases,"* said **Darrell Phillips, vice president of human re-**

> **sources for Hollingsworth & Vose**. *"But you have to recognise that these events took place 35 years ago. It has no bearing on our work force today."* (wsj_0003)

Each attribution relation was reconstructed, further annotated, as described in Section 3.3.2, and stored as stand–off CoNNL annotation. The annotation includes, for each attribution, columns corresponding to the elements showed in Table 3.5, together with byte references to the original text for each annotated span.

```
____Explicit____
3904..3907
#### Text ####
But
#### Features ####
Ot, Comm, Null, Null
3820..3901
#### Text ####
said Darrell Phillips, vice president of
human resources for Hollingsworth &
Vose
####but, Comparison.Contrast
____Arg1____
3721..3817
#### Text ####
There's no question that some of those
workers and managers contracted
asbestos-related diseases
#### Features ####
Inh, Null, Null, Null
____Arg2____
3908..3971
#### Text ####
you have to recognize that these events
took place 35 years ago
#### Features ####
Inh, Null, Null, Null
```

```
____Implicit____
3973
#### Features ####
Ot, Comm, Null, Null
3820..3901
#### Text ####
said Darrell Phillips, vice president of
human
resources for Hollingsworth & Vose
####in other words, Expansion,
Contingency
____Arg1____
3930..3971
#### Text ####
that these events took place 35 years
ago
#### Features ####
Ot, Ftv, Null, Indet
3908..3929
#### Text ####
you have to recognize
____Arg2____
3973..4014
#### Text ####
It has no bearing on our work force
today
#### Features ####
Inh, Null, Null, Null
```

Figure 3.1: Annotation of attribution in the original release of the PDTB 2.0 (Prasad et al., 2008). Each column reports the annotation relative to a discourse connective and its arguments, including its attribution.

3.3.2 Further Annotation

The collected ARs were further annotated in order to distinguish the elements in the 'attribution span'. In the PDTB annotation the attribution span includes the source as

ATTRIBUTION ID:	wsj_0003.pdtb_05
SOURCE SPAN:	Darrell Phillips, vice president of human resources for Hollingsworth & Vose
CUE SPAN:	said
CONTENT SPAN:	“There’s no question that some of those workers and managers contracted asbestos–related diseases,”|“But you have to recognise that these events took place 35 years ago. It has no bearing on our work force today.”
SUPPLEMENT SPAN:	None
FEATURES:	Other, Assertion, Null, Null
QUOTATION TYPE:	Direct

Table 3.5: Example of an AR in the initial version of PARC.

Rule	Example
(NP-SBJ)(VP)	**one person** said
(PP-LOC) **(NP)**(VB)	IN DALLAS, **LTV** said
(NP-SBJ)(VBP)(JJ)	**I** am sure

Table 3.6: Examples of patterns for the fine-grained annotation of the PDTB reporting spans into **source** cue and SUPPLEMENT spans.

well as the cue spans and additional elements. Within the attribution span, the spans corresponding to source and cue had to be identified while the remaining text could be marked as supplement if considered relevant to the AR. Around 80% of the annotation was performed semi–automatically by making use of a system of 48 syntactic rules such as the ones in Table 3.6, to identify the most common source–cue patterns. The identified spans were then manually revised. The remaining 20% of attribution spans presented less common structures, thus requiring manual annotation. Both revision and annotation were performed by one expert annotator.

Elements of the attribution span were marked as **source**, cue or SUPPLEMENT, according to the annotation schema developed in Pareti and Prodanof (2010) and described in Section 3.1. The source comprises the source mention together with its description, usually in the form of an appositive as in Ex. (65) or a relative clause. In case of a source expressed by a possessive adjective as in Ex. (66) or pronoun, the whole NP was annotated.

(65) **Pierre-Karl Peladeau, the founder's son and the executive in charge of the acquisition,** says *Quebecor hasn't decided how it will finance its share of the purchase*, but **he** says *it most likely will use debt.* (wsj_0467)

(66) **His point**: *It will be increasingly difficult for the U.S. to cling to command-and-control measures if even the East Bloc steps to a different drummer.* (wsj_1284)

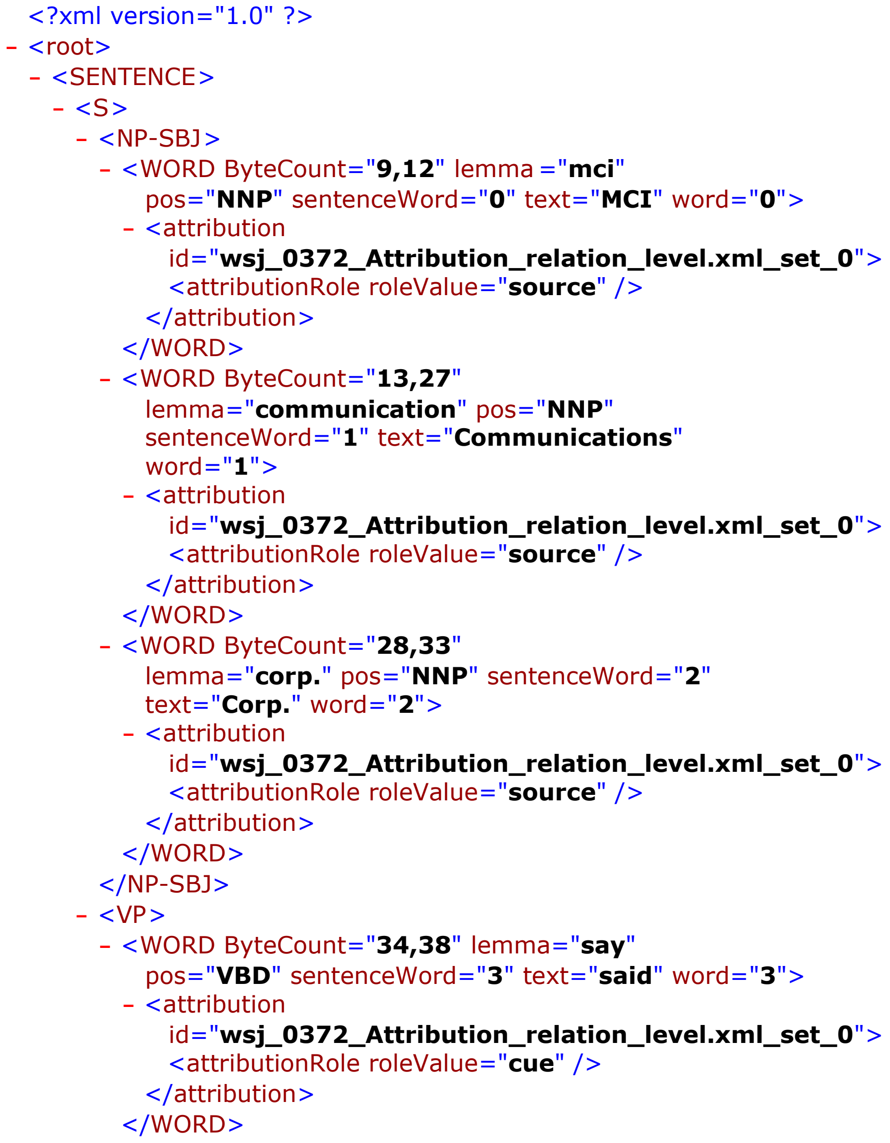
Verbal cues were annotated together with their full verbal group, including auxiliaries, modals and negative particles. Adverbials adjacent to the cue, as in Ex. (67), were also included, since they can modify the verb. Other parts of the verbal phrase were marked as supplement. Prepositional cues (e.g. 'according to', 'for'), adverbial cues (e.g. 'supposedly', 'allegedly'), and noun cues (e.g. 'pledge', 'advice') were also annotated.

(67) *"I'm not sure he's explained everything,"* **Mrs. Stinnett** says grudgingly. (wsj_0413)

All additional elements within the attribution span that were relevant for the interpretation of the content, but not strictly part of the attribution were annotated as supplement. This includes circumstantial information, such as time (e.g. 'People familiar with Hilton said OVER THE WEEKEND' (wsj_2443)), location, manner, topic (e.g. 'ON THE PROVISIONS OF THE MINNESOTA LAW, the Bush administration said ...' (wsj_2449)) and recipient ('He told THE WOMAN'S LAWYER, VICTOR BLAINE ...' (wsj_0469)). Punctuation was also added to the attribution corpus in order to distinguish between direct, indirect and mixed attributions.

The first PARC version (PARC 1.0) comprises 9,868 ARs collected and further annotated from the PDTB annotation. ARs having a discontinuous attribution span were not included in this version. PARC 1.0 is in a CoNLL-like tabular style format with stand-off annotation and was employed for the preliminary analysis of ARs.

Subsequently, ARs with a discontinuous attribution span were revised and also included in the corpus. The corpus also underwent a preliminary revision of incomplete or incorrect ARs. The final version comprises 9,893 ARs and is identified as PARC 2.0. PARC 2.0 annotation is in-line and was added to the PTB merged files, which comprise POS and syntactic annotation after converting the bracketed annotation into an XML tree. Nodes represent syntactic nodes as well as terminal words. Attribution nodes were added as children for each token part of an AR as shown in Fig. 3.2. A version of PARC 2.0 including only ARs of type *assertion* was used for most of the preliminary experiments described in Ch. 4.



```
  <?xml version="1.0" ?>
- <root>
  - <SENTENCE>
    - <S>
      - <NP-SBJ>
        - <WORD ByteCount="9,12" lemma ="mci"
            pos="NNP" sentenceWord="0" text="MCI" word="0">
          - <attribution
              id="wsj_0372_Attribution_relation_level.xml_set_0">
              <attributionRole roleValue="source" />
            </attribution>
          </WORD>
        - <WORD ByteCount="13,27"
            lemma="communication" pos="NNP"
            sentenceWord="1" text="Communications"
            word="1">
          - <attribution
              id="wsj_0372_Attribution_relation_level.xml_set_0">
              <attributionRole roleValue="source" />
            </attribution>
          </WORD>
        - <WORD ByteCount="28,33"
            lemma="corp." pos="NNP" sentenceWord="2"
            text="Corp." word="2">
          - <attribution
              id="wsj_0372_Attribution_relation_level.xml_set_0">
              <attributionRole roleValue="source" />
            </attribution>
          </WORD>
        </NP-SBJ>
      - <VP>
        - <WORD ByteCount="34,38" lemma="say"
            pos="VBD" sentenceWord="3" text="said" word="3">
          - <attribution
              id="wsj_0372_Attribution_relation_level.xml_set_0">
              <attributionRole roleValue="cue" />
            </attribution>
          </WORD>
```

Figure 3.2: PARC 2.0 XML annotation format.

3.4 Final Version

In this section, I describe the annotation effort that was undertaken in order to create PARC 3.0, a large and complete corpus annotated with ARs. The corpus aims at providing a rich basis for attribution studies. Although already a large resource for attribution, not all ARs are annotated in the early versions of PARC. Any analysis based on the incomplete data is thus only tentative as it presupposes the annotated ARs being a representative and balanced subset of all ARs in the corpus. However, this is not the case since the annotation is subordinate and dependent on that of discourse relations.

In addition, incomplete data is also detrimental for the development of supervised attribution extraction components which are confronted with the challenge of learning from positive instances and unlabelled data. While it is possible to overcome this issue, having a completely annotated resource is preferable.

The initial corpus was therefore further annotated with missing and nested ARs. The resulting corpus, PARC 3.0, includes 19,712 ARs and is divided into three sections corresponding to the WSJ corpus folders:

- Train: folders 00-22
- Development: folder 24
- Test: folder 23

The annotations originate from three distinct annotation phases:

1. PDTB derived: around half of the ARs are derived from the partial annotation in the PDTB. They were reconstructed and their ‘attribution span’ further annotated as ‘source’ and ‘cue’ as described in Sec. 3.3. There are some annotation errors in the original annotation, in particular some incomplete content spans. These have not been corrected.
2. New annotation: annotation of all missing first-level ARs
3. Nested annotation: annotation of nested ARs in the development and test sections and folders 0-11 of the training section.

New annotations of first-level and nested ARs were added only to the 1,833 WSJ documents classified as news [3]. *News* is by far the largest genre in the WSJ corpus. The following section will describe the annotation work.

Like PARC 2.0, PARC 3.0 annotation is in-line and encoded in XML. Tokens that are part of an annotation have an *attribution* child element, containing the AR unique id and one or more *attributionRole* children as in the examples below:

[3] A list of WSJ documents per genre: `http://www.let.rug.nl/~bplank/metadata/genre_files_updated.html`

(68) <WORD ByteCount=”1717,1721” gorn=”10,3,1,5,1,3,1,1,1,1,1,1,0,2”

lemma=”tune” pos=”NN” sentenceWord=”56” text=”tune” word=”330”>

<attribution **id=wsj_0207_PDTB_annotation_level.xml_set_0**>

<attributionRole roleValue=”content” />

</attribution>

</WORD>

The unique AR identifier **id** specifies:

- wsj file name:wsj_0207
- annotation origin: either 1)PDTB or 2)Attribution (i.e. new first-level annotations) or 3)Nested
- set number: set_N (all elements belonging to the same AR are grouped in a set)

3.4.1 Annotation

The new annotation was manually performed by three linguist annotators that worked part-time over a period of four months. The annotators underwent an initial training phase in order to familiarize themselves with the task and the annotation guidelines. In this phase the annotators independently annotated the same texts and were then asked to jointly review and discuss any disagreement. Subsequently, the annotators proceeded independently but were able to discuss with the other annotators and the instructor in case of uncertainty. Doubts and borderline cases were collected during the annotation and the annotators and instructor met regularly to discuss them and incorporate some clarifications into the guidelines.

The full guidelines are included in Appendix B and are a modification of the guidelines in Appendix A adopted for the preliminary inter-annotator agreement study described in Section 3.2. The annotation was simplified by removing the selection of the attribution features and by decoupling the annotation of first-level and nested ARs. The annotators were first asked to annotate first-level ARs (Task 1) and once that task was concluded for all texts, two annotators proceeded with the annotation of nested ARs (Task 2). Potential verb-cues were automatically identified using the supervised classifier developed on PARC 2.0 and described in Section 4.2.2. These verbs were

Folder	Texts
00	00-99
01	00-59
03	00-08
10	00-11

Table 3.7: Double annotated PARC 3.0 texts.

shown to the annotators as highlighted in the text in order to attract their attention on likely ARs.

For Task 1, text that belonged to an already annotated AR from PARC 2.0 was greyed out not to distract the annotators and to reduce the annotation time. Since annotators were not revising PDTB-derived annotations and did not have to consider nested ARs at this stage, the existing annotation could be safely ignored.

For Task 2, all text was greyed out with the exception of the content spans of all annotated ARs, whether coming from PARC 2.0 existing annotation or being the ones added in Task 1. For this task, the annotators had to solely consider nested ARs. Since nested ARs occur inside the content of another AR, they where required to examine only the text portion corresponding to previously identified content spans.

The texts were single-annotated, apart from a subset of approximately 7% of the texts, listed in Table 3.7, that were double annotated. Double annotated texts were part of the initial training or used to monitor the inter-annotator agreement. Disagreement on these texts was then adjudicated by a third annotator. The results of the inter-annotator agreement study are presented in the following section.

3.4.2 Inter-annotator Agreement

Approximately 7% of PARC 3.0 news texts were double-annotated, which allowed me to compute reliable inter-annotator agreement scores for the identification of ARs and for the selection of the spans corresponding to source, cue, content and supplement.

A large proportion of the double-annotated texts are part of the training phase, while a smaller number of texts was double-annotated at certain intervals in order to monitor the consistency and quality of the annotation.

Table 3.8 reports the overall *agr* results for the double-annotated texts, including the initial texts that were still part of the training phase. For the identification of an AR, the *agr* for each annotators pair varies from .74 to .82, while the overall *agr* is .79.

Annotators	Texts	ARs Ann1	ARs Ann2	ARs both	Agr
AB	39	311	266	235	0.82
BC	38	166	128	116	0.80
AC	42	229	266	183	0.74
Overall	119	706	660	534	0.79

Table 3.8: Inter-annotator agreement results for the new annotation in PARC 3.0. Texts were two-fold annotated and scores for the identification of an AR are presented for each annotator pair and averaged for all annotators.

When considering only texts that were double annotated after the training was complete, the *agr* score is .83 (Table 3.9). The score is considerably lower than the .87 agr score reported in Sec. 3.2 for the identification of ARs in the preliminary inter-annotator agreement study. That annotation task was more complex, the texts were selected to be long and particularly rich in ARs and the training was less thorough.

However, the annotators from the first agreement study were experienced annotators familiar with the task and they annotated all ARs in the text, while in the second study, annotators added the missing ARs. Thus the agreement score refers not to the identification of all ARs, but of those that were not picked up by the PDTB annotation. PDTB derived ARs may be more prototypical and thus their identification less problematic.

Annotators	Texts	ARs Ann1	ARs Ann2	ARs both	Agr
AB	4	30	28	26	0.90
BC	12	45	43	37	0.84
AC	18	83	103	70	0.76
Overall	34	158	174	133	0.83

Table 3.9: Inter-annotator agreement results for the new annotation in PARC 3.0, excluding texts annotated during the trainig phase. Texts were double annotated and scores for the identification of an AR are presented for each annotator pair and averaged for all annotators.

Another reason for the difference in score derives from the high variation of complexity from text to text. While the first agreement study included only 14 articles, the second one comprises a larger variety of texts and in particular legal and economics news texts which presented recurrent problematic cases due to domain specific characteristics (e.g. orders and court decisions, laws) and terminology. In Ex. (69), all verbs in bold are likely AR cues, however they are part of a rather fixed legal expression.

(69) Without **admitting** or **denying** wrongdoing, they **consented** to findings of violations of escrow and record-keeping rules. Mr. Crane didn't return a call seeking comment.(wsj_0096)

Fig. 3.3 reports the AR identification *agr* fo each individual PARC 3.0 text that was double annotated. While texts that are extremely rich in ARs are more complex and have a lower *agr*, but a more stable score, for the rest of the texts, *agr* varies extremely, even for texts with a similar amount of ARs.

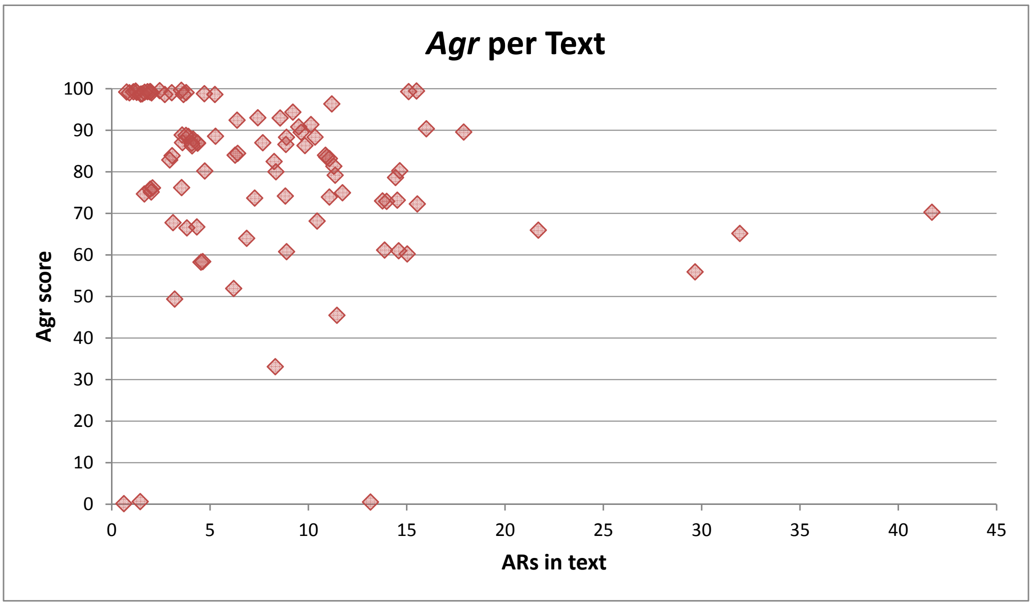


Figure 3.3: *Agr* score for the identification of an AR per text, considering the overall number of ARs identified in the text.

For the commonly identified ARs it is possible to compute the *agr* for the annotation of the spans corresponding to *source*, *cue*, *content* and *supplement*. Overlap results are calculated by taking the mean of the *agr* scores for each individual span. The results, reported in Table 3.10, are very encouraging, with cues being almost always commonly identified with exact boundaries and source and content spans having also very high *agr*: .91 and .94 respectively.

Since for a large proportion of ARs no supplement was identified, the *agr* for the supplement span was calculated by taking into account only the ARs for which a supplement was identified. The score of .46 *agr* is rather low. However, the annotation of the supplement was included as exploratory of the kind of elements that would also be relevant for an AR. The annotation of one or more supplemental element was left optional and underspecified in order to learn from the annotation instead of forcing it into a predefined direction.

Annotators	ARs	Cue	Source	Content	Supplement
AB	26	1.00	0.90	0.95	0.67
BC	37	1.00	0.94	0.94	0.50
AC	70	0.99	0.88	0.95	0.30
Overall	133	1.00	0.91	0.94	0.46

Table 3.10: PARC 3.0 span selection overlap *agr* metrics for each annotation pair and averaged to calculate the overall agreement.

The inter-annotator agreement was also calculated, on a small set of texts, for the task of annotating nested ARs. The overall *agr* for this task was .70 (Table 3.11). It is possible to argue that nested ARs are more complex and less prototypical, since they are almost never signalled by quotation marks, are expressed by different and shorter structures and contain a larger proportion of beliefs and eventualities, which are a harder set of ARs with respect to assertions. However, the smaller size of the agreement study makes this result only indicative, since a different subset of texts would likely determine rather different scores.

Annotators	Texts	ARs Ann1	ARs Ann2	ARs both	Agr
AB	11	29	38	23	0.70

Table 3.11: Inter-annotator agreement results for the new annotation of nested ARs in PARC 3.0.

The results of the agreement study reveal the heterogeneous and pervasive nature of attribution. While a number of relations can somewhat entail attribution, some are more prototypically associated with attribution and others are more borderline. In particular, by looking at the disagreement, it emerges that assertions are more clearly associated with attribution than beliefs, facts or eventualities. Similarly, more common and standard sources and cues contribute to identifying the AR. ARs having a finite clausal content are also more easily identified with respect to those whose content is a non-finite clause, a phrase, or anaphoric, while non-factual ARs cause higher disagreement.

Although satisfactory, agreement scores for the identification of ARs seem to set a relatively low ceiling for attribution extraction systems. However, we should consider that these scores do not include the annotation of ARs that were already in PARC 2.0, which may be expressed by more identifiable structures due to the constraints set by the PDTB annotation. Namely, the overlap with discourse connective arguments causes ARs having a finite clausal contents to be overrepresented. These associate more with assertions and are more unanimously identified by the annotators.

In conclusion, the annotation reliably identifies those ARs that are more standard and more strongly associated with their attributive function with high agreement. There exists a number of cases where the attributive function is perceived as less clear or less relevant to identify for which disagreement is higher.

3.5 Attribution Relation Analysis

The analysis presents results on the whole PARC 3.0 corpus, which includes PDTB derived ARs as well as the new first-level annotations and the nested ARs. Statistics and analysis concerning nested ARs are calculated only relative to texts fully annotated also with nested ARs (i.e. folders 0-11, 23 and 24).

3.5.1 Source

The source is explicitly expressed in 92% of the ARs. The remaining are cases where a passive structure, an adverbial cue (e.g. 'reportedly') or ellipsis of the subject in a coordinate or subordinate clause conceal the source.

Excluding the implicit sources, source spans are 3.7 tokens long on average. However, complex noun modifiers are relatively common and source spans including appositives, prepositional arguments and relative clauses might be longer than their content span and occasionally exceed 35 tokens as in Ex. (70).

(70) *"You have to go out to all your constituents,"* says **James H. Giffen, who is spearheading the most ambitious attempt by U.S. firms to break into the Soviet market, involving investment of more than $5 billion in some two dozen joint ventures**. (wsj_1368)

The vast majority of source spans consist of noun phrases, and over 83% of AR sources are noun phrases in subject position.

Concerning the common assumptions that sources correspond to named entities (NE), the annotations from the corpus downsized their importance. Although proper nouns are a relative majority of sources (40%) as reported in Table 3.12, a considerable number of them are expressed by common nouns (30.7%), only in part referring to an NE. In particular, plural common nouns (e.g. 'lawyers', 'officials', 'people', 'nerds', 'libertarians' and 'enthusiasts') usually refer to categories of people and hardly ever to NEs. Another common type of sources is represented by pronouns (personal (19.3%) but also relative or who (1.3%), indefinite and demonstrative (0.9%) and pronominal cardinal numbers (0.2%), some of which will refer to NEs, while others not. In addition, 7.7% of ARs have an implicit source. This can correspond to a precise or a generic entity.

Element	Occurrence	Percentage	Examples
NE	7894	40.0	*Bowder, Fed Chairman Greenspan*
noun	6053	30.7	*an official, analysts, most people*
pronoun	3800	19.3	*they, his, I*
implicit	1510	7.7	NONE
wh. pronoun	250	1.3	*who, which, that*
determiner	173	0.5	*some, many at Lloyd's*
numeral	32	0.2	*the two, one in ten*

Table 3.12: Type of AR sources in PARC 3.0 (occurrence and percentage).

3.5.2 Cue

Verbs are by large the most frequent type of cues in the corpus, covering 92% of the cases. The remaining 8% is represented by a range of different elements as summarized in Table 3.13. Almost 4% of cues are nouns and another 2% are prepositional groups, almost exclusively "according to". The remaining types are relatively infrequent and are: adjective cues in copula construction; prepositions; punctuation markers only and adverbials and possessives.

Only 22% of the cue spans are longer than a single token. These include those corresponding to prepositional groups, verbs including adverbial modifiers, negations and auxiliaries and adjective modifiers.

Element	Occurrence	Percentage	Examples
verb	18136	92	*say, want, shrug*
noun	765	3.9	*announcement, idea, word*
according to	392	2.0	according to, in the eyes of
adjective	244	1.2	*is sure/skittish/aware*
preposition	81	0.4	*under, for, by, in, to*
punctuation	50	0.3	*colon, quotation*
adverbial	34	0.2	*admittedly, unexpectedly, reportedly*
possessive	1	0.0	*'s (Mr. Mushkat's "realists")*

Table 3.13: Type of attributional cue in PARC 3.0 (occurrence and percentage).

3.5.2.1 Verbs

Verbs are not the only possible AR cues and should not be the only type of cue taken into account. Nonetheless, they deserve particular attention as they represent the most common AR textual anchor (92%).

There are 527 different attributional verbs in the corpus. The top 20 most frequent verbs are reported in Table 3.14 and the full list of the attributional verbs in PARC 3.0 is in Appendix D. While the number of verb types is large, their occurrence is strongly skewed. On the one hand, 'say' alone accounts for approximately half (49.7%) of the occurrences of a verb-cue, on the other, 40% of the verb types (199 types) are hapax legomena and a similar number of them have low occurrence as an AR cue (2-9 occurrences). The top 50 verbs cover around 83% of the occurrences.

The verb type distribution suggests that using small lists of attributional verbs can be relatively effective but would still miss a relevant proportion of attribution cues. Compiling a more comprehensive list would struggle to capture the long tail of verb types that are rarely used or can occasionally assume an attributional meaning. Even among the most frequent 50 verbs, there are several that are common verbs, whose attributional use or meaning gets activated only in specific contexts, such as 'add', 'show' and 'find'.

Top frequency verb-cues are more neutral as they usually adhere to the general principle according to which the journalist should report the facts but remain neutral. In lower frequency verb-cues, however, the reporter's voice is mostly expressed. By choosing a non standard verb, the journalist already makes a marked choice. Verbs in the lower end of the frequency scale are:

Lemma	Occurrence	Percentage	Lemma	Occurrence	Percentage
say	9017	49.7	announce	186	1.0
expect	671	3.7	plan	175	1.0
add	372	2.1	consider	131	0.7
think	333	1.8	estimate	130	0.7
report	313	1.7	know	128	0.7
believe	267	1.5	ask	127	0.7
want	253	1.4	call	122	0.7
note	241	1.3	argue	121	0.7
agree	233	1.3	predict	101	0.6
tell	191	1.1	cite	95	0.5

Table 3.14: Top 20 most frequent verb-cues in PARC 3.0 (occurrence as attributonal and percentage over all verb-cues).

- attack, castigate, chide, feud, erupt, frighten, fume, wrestle
- bemoan, grouse, grumble, irk
- chuckle, croon, crow, flirt, gloat, muse, prim, rave, trumpet
- couch, harp
- marvel, caricature

Also in case of a relatively neutral verb-cue choice, modifiers can contribute to a connotated meaning. Adverbial modifiers are not extremely common, however they are worth including as they can affect the verb-cue by expressing the authorial stance (e.g. ‘optimistically’, ‘unrealistically’), the manner (e.g. ‘emphatically’, ‘solemnly’, ‘darkly’, ‘forcefully’), the attitude of the source (e.g. ‘proudly’, ‘sardonically’, ‘derisively’, ‘apologetically’, ‘grudgingly’), circumstantial information (e.g. ‘privately’, ‘recently’) and can even change its factuality (e.g. ‘not’).

3.5.2.2 Other Cues and Special Cases

Ignoring cues other than verbs would ignore 8% of ARs in the corpus. Most non-verbal AR cues are nouns often closely related to an attributional verb, e.g.:

- to think - thought

- to announce - announcement
- to fear - fear
- to claim - claim
- to agree - agreement

Sources of noun cues are implicit in 40% of the cases. The rest of the cases tend to either have the source as a prepositional dependent (e.g. ‘a statement by Mr. Keating’, ‘a general proposal from State West’) or as a possessor (e.g. ‘his advice’). Occasionally, the noun cue is part of a verbal construction (e.g. ‘had the idea’ and ‘made a statement’). Noun cues can be negated with a negative particle, preposition or adjunct (e.g. ‘had no plans’, ‘without fears’, ‘with nary a mention’), but also with a negative prefix on the noun itself (e.g. ‘unwillingness’).

Adjectival cues usually take the form of a copula construction as in Ex. (71), with the source in subject position. Alternatively, the complete AR can be part of a noun phrase, with the source as the head noun and the content span as a clausal complement of the adjectival cue as in Ex. (72).

(71) **Sen. Mitchell** is confident *he has sufficient votes to block such a measure with procedural actions.* (wsj_0343)

(72) **People** eager *to have youth “pay their dues to society”* favor service proposals – preferably mandatory ones. (wsj_2412)

Cues can also be expressed by a simple preposition, as in Ex. (73), where the cue is a prepositional modifier of the content and the source is the object of the preposition.

(73) But by **most accounts**, *he made little of the post and was best known among city politicians for his problems making up his mind on matters before the city’s Board of Estimate*, the body that votes on crucial budget and land-use matters. (wsj_0765)

In infrequent cases, there is no textual cue other than punctuation that suggests the presence of an AR. In those cases the colon introducing the content clause or the quotation marks surrounding it are considered the cue as in Ex. (74).

(74) **Mr. Rogers** spent half his cash on hand Friday for *“our favorite stocks that have fallen apart.”* (wsj_2381)

Also infrequent are adverbial cues as in Ex. (75). These express the AR by evoking an implicit third-party source that is not better specified.

(75) *Olivetti* reportedly *began shipping these tools in 1984 (wsj_2326).*

An anomaly that could be observed in some rare cases in the corpus is the use of redundant cues, leading to what could be called a "two-headed" content span. In Ex. (76), there is one content span crossing sentence boundaries. Since the boundary is within the quotation marks, the content span is not split into two. At the end of the span in the second sentence, however, the author anaphorically recalls the source and repeats the cue as if to refresh them in the reader's mind.

(76) Though the ink is barely dry on its new, post-bankruptcy law structure, **Bill Bullock, Manville's head of investor relations**, says *the company is continually pondering ”whether there is a better way to be structured. We understand that the trust is ultimately going to need to sell some of our shares,”* **he** says. (wsj_1328)

In the corpus there are also six poems that were incorrectly annotated as an AR. While there is an AR connecting the poem to its author, the relation is not within the text but rather meta-textual similar to the one between the news text and its author, which is left implicit. In such cases there is no textual cue expressing the relation, since this is inferred by our knowledge of what a poem is and the position where we could expect to find the name of the author

(77) Rex Tremendae/ The effete Tyrannosaurus Rex/ Had strict Cretaceous views on sex,/ And that is why you only see him/ Reproduced in the museum./ – Laurence W. Thomas. (wsj_1758)

3.5.3 Content

Any span of text can potentially be the content span of an AR. Contents in PARC 3.0 are between 1 and 500 tokens long (Fig. 3.4), with an average of 19.6 tokens for first-level ARs and 11.5 for nested ones. While ARs are mostly identified at the intra-sentential level, the relation can cross sentence boundaries. The data contains 1,727 ARs spanning over 2 to 27 sentences, as the one in Ex. (78) which comprises several paragraphs corresponding to a list from a political proposal. Around 12% of AR contents are discontinuous. This is usually the case when the attribution span expressing the source and cue is in a parenthetical construction.

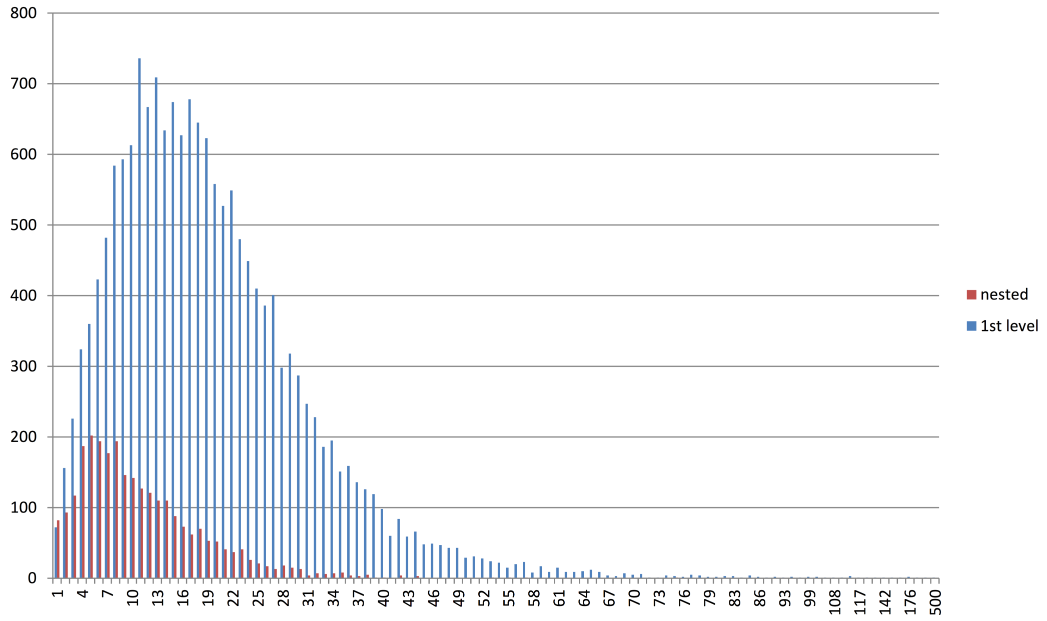


Figure 3.4: Content length in tokens for first-level (blue) and nested (red) ARs in the PDTB.

(78) The key steps advocated include:

– PROPERTY. Rigid ideological restrictions on property ownership should be abandoned.[...]

– FOREIGN TRADE. The current liberalization and decentralization of foreign trade would be taken much further.[...] (wsj_0756)

Contents can be expressed by virtually any syntactic structure, however most content spans correspond to a clausal element. This is an SBAR clause, i.e. a clause introduced by a subordinating conjunction in over 37% of the cases and in particular for ARs with the content following the source and cue spans. When the content span precedes the source and cue spans, it mostly consists of a declarative clause (S) or a topicalized declarative clause (S-TPC), i.e. the clause is before the subject. Around 24% of AR contents fall in this group. Also relatively frequent, around 8% of the cases, is the content corresponding to one or more noun phrases (NP) as in Ex. (79). The remaining cases are often a combination of S/SBAR/NP and other structures, such as a complete sentence, another clause or a phrase.

(79) Even if the **government** does see *various "unmet needs,"* national service is not the way to meet them. (wsj_2407)

3.5.4 Features

3.5.4.1 Level of Nesting

Nested ARs are almost absent from the literature and their extraction has yet to be addressed, nonetheless, they are rather frequent, particularly in news. In PARC 3.0, there are 2,689 nested ARs annotated. However, to correctly quantify their incidence, we have to consider only those texts that were annotated with nested ARs (the second stage of annotation on PARC 3.0 as described in Sec. 3.4.1). On the fully annotated texts corresponding to the news texts in folders 00-11, 23 and 24 of the WSJ corpus, the percentage of nested ARs is over 20% as Table 3.15 reports. This translates to almost 1 in 4 first-level ARs carrying a nested AR within their content span.

Level of Nesting	All Texts	Nested-annotated Texts
1st	17016 (86.4)	9747 (79.7)
2nd	2526 (12.8)	2321 (19.0)
3rd	163 (0.8)	161 (1.3)

Table 3.15: Level of Nesting distribution in PARC 3.0. Occurrence (and percentage) of first-level and nested (2nd and 3rd level) ARs. Results are given for the complete PARC 3.0 and relative to the texts that were specifically annotated with nested ARs (i.e. news texts in folders 0-11, 23 and 24 in the corpus).

Nesting can be thought of as a distance measure, or the path the information went through to reach the text we are reading. Thus first-level ARs are just one step away from the author of the text and imply only one additional source, while nested ones went through two or more passages before reaching the text. While theoretically it is possible to reach a deep level of nesting, this is rather infrequent. In PARC 3.0, most nested ARs are second-level, while 6.5% of them are third-level (as in Ex. (80) and (81)).

(80) Lately, **analysts** say, $_{\text{1st}}$[Deutsche Bank has shocked some in the French financial community by indicating $_{\text{2nd}}$[**it** wants $_{\text{3rd}}$[*a strong bank with a large number of branches*]]]. (wsj_0477)

(81) **The company’s prepared statement** quoted $_{\text{1st}}$[**him** as saying, $_{\text{2nd}}$[“The CEO succession is well along and **I** $_{\text{3rd}}$[’ve decided for personal reasons *to take early retirement*.”]]] (wsj_0109)

3.5.4.2 Quote Status

Direct, indirect and mixed ARs present various characteristics and complexity. While the content span of a direct AR is easily identified, that of a mixed AR has less clear boundaries and that of an indirect AR cannot be identified based on punctuation clues. Hence, the quote status of an attribution affects the complexity of the annotation and the success of an AR extraction system.

While the main focus of attribution extraction studies is on direct ARs, the quote status distribution in PARC 3.0, presented in Table 3.16, downsizes their relevance. Direct ARs account for just over 14% of all ARs, the same portion also corresponds to mixed, while 72% are indirect ARs. There is also a significant difference in distribution between nested and non-nested ARs. For nested ARs, the percentage of direct ones drops to just 1.7% and that of mixed to 10.6%.

Quote Status	Non-nested	Nested	All
Direct	2771 (16.2)	45 (1.7)	2816 (14.3)
Indirect	11823 (69.3)	2361 (87.6)	14184 (72.0)
Mixed	2464 (14.4)	286 (10.6)	2750 (14.0)

Table 3.16: Quote Status distribution in PARC 3.0. Occurrence (and percentage) of direct, indirect and mixed ARs.

Nested ARs are in fact mostly indirect, since direct reporting presupposes a verbatim of the original utterance, which becomes less likely, and credible, for nested ARs. In Ex. (82), the nested AR content appears as direct, however, it is unclear whether the first source is reporting the exact words of the nested source or rather the quoted portion is a verbatim of the spokesman's description of what the nested source feels.

(82) A spokesman for Rep. Edward J. Markey (D-Mass.), who heads a subcommittee that oversees the FCC, says **Mr. Markey** feels *"the world has been forever changed by the Sony-Columbia deal."* (wsj_2451)

3.6 Conclusion

The creation of a new corpus annotated with discourse relation was presented in this chapter. The annotation scheme is a modification of the PDTB annotation scheme for attribution. The initially proposed scheme included three constitutive elements, i.e. the source, the cue and the content as well as an optional one, the supplement, and a set of features: attribution type, authorial stance, source attitude, source type, factuality, scopal polarity, quote status and level of nesting.

With respect to the PDTB annotation scheme, the modified scheme (partly developed in Pareti and Prodanof (2010)) further classifies the 'attribution span' into source and cue and introduces the supplement as a generic label for additional information that affects the AR, e.g. recipient or circumstantial information. Concerning the feature set, authorial stance and source type were added to the four types already in the PDTB, together with two automatically derived one: quote status and level of nesting.

Quote status identifies whether the content of an AR is a direct, indirect or mixed quotation. All facts, beliefs and eventualities are always indirectly reported.

'Level of nesting' accounts for the depth of an attribution, i.e. the AR is nested into another AR, and as such is also a measure of reliability. Not only since the information conveyed in the AR content is second or third-hand (or more), but also because there are more sources involved and their bias and credibility will affect whether we trust the AR they establish to be truthful and the conveyed information to be accurate. For each AR, the 'level of nesting' can be reliably computed by counting the number of AR contents it is contained within, taking the text as the zero level.

The scheme was tested by conducting an inter-annotator agreement study on a set of 14 articles. The results showed a relatively high agreement for the identification of an AR (*agr* .87) and a high agreement for the selection of each constitutive element span (*agr*: .97 cue, .94 source, .95 content). However, the results also highlighted some flaws in the scheme and the necessity to break the annotation task into more manageable steps. In particular, the agreement for the features was mostly not satisfactory, in part because the proposed categories were imbalanced and appeared problematic to identify. While the features would be a valid addition to a corpus of ARs, further investigation would be required. None of the manual features were therefore included in further annotations.

A first corpus of over 9,800 ARs, PARC 1.0, was compiled from existing PDTB annotations that were reconstructed and further annotated semi-automatically. This version was used to conduct preliminary analysis of attribution. After some revision and correction work on PARC 1.0, PARC 2.0 was completed and employed in the first experiments on the automatic extraction of ARs.

Since a major drawback of the preliminary PARC versions was the data being only partially annotated, a second round of annotation was conducted in order to have a complete resource. This lead to PARC 3.0, a corpus of almost 20k ARs. The corpus has enabled studying ARs and identifying how they are expressed. In particular, the analysis could identify the large proportion of nested ARs (1 in 5). It has also confirmed that a large proportion of ARs sources are not named (only 40% are expressed by a proper noun). Concerning the cue, PARC 3.0 contains 527 attributional verb types, 40% of which occur a single time as an AR cue, thus relying on a pre-compiled list of verbs for attribution extraction is not a satisfactory solution. Moreover, while cues are mostly verbs, in 8% of ARs the cue is not a verb, thus focussing on verbs only would miss those relations.

Chapter 4

The Automatic Extraction of Attribution Relations

Different studies have addressed the extraction of a subset of ARs, e.g quotations or opinions, or a portion of the relation, e.g. the attribution of the content span to its source. However, there is no trace in the literature of complete attribution relations (ARs) of different types being automatically extracted. For example, assertion and belief attributions have only been tackled in separate studies (see Sec. 2.2.1 and Sec. 2.2.3). Concerning the AR components, the identification of AR cues has been widely neglected.

This chapters describes a methodology for the automatic extraction of all types of ARs found in the PDTB. (I briefly mention the problem of extending this work to cover broad AR extraction from other genres under Ch. 6, Future Work). The methodology consists of a pipeline of models implementing a sequence of steps to identify and link *source*, *content* and *cue* spans of each AR.

Sections 4.2 and 4.3.1, in part, already appeared in Pareti et al. (2013), while Section 4.4 has appeared in part in O'Keefe et al. (2012). The contribution of the author to sections derived from joint work is further clarified in each related section.

4.1 Methodology

The complex task of extracting ARs is addressed in this chapter as a sequence of smaller tasks joined in a pipeline model. Each subtask will be presented separately and their output and results recombined. The model architecture is shown in Fig. 4.1 and each corresponding step and component of the system is discussed in Sec. 4.1.2.

The pipeline model starts with the identification of attribution cues. This was chosen as the starting point for three main reasons:

- Cues function as the element that establishes the AR. Their identification is thus a strong indicator of the presence of an AR.
- In the current approach, attribution cues are lexically anchored and unique, therefore for each AR there is one and only one attribution cue and this is expressed as a text span. In contrast, an AR might have implicit or multiple sources and separate spans or an anaphoric pronoun corresponding to its content.
- Cues are predominantly expressed by verbs and a relatively small set of them will cover the majority of the cases. Sources instead can be proper or common nouns, pronouns and also complex noun phrases while any text span can be a content span. It is therefore a less complex task to identify the majority of AR cues.

Moreover, sources and contents are usually identified by their relation with the cue and it is therefore critical to be able to identify potential cues first. Instead of looking for content and source spans for each identified cue, cues are used to generate features that can help the identification of source and content spans.

After the identification of potential cues, the system tackles the extraction of the content span. This step is the most complex and therefore most prone to error and it would seem best to address it towards the end of a pipeline system in order to reduce the amount of errors that get propagated. However, we first need to identify the content span in order to rule out nested ARs, which are not addressed by the current model. Once first-level (i.e. non-nested) ARs content spans are identified, sources and cues within it are no longer taken into account. Starting from the identification of source and cue spans instead, we would also identify sources and cues related to nested ARs which would compromise the correct identification of the content span.

Being interconnected, the identification of each of the three components of the AR would benefit from having already identified the other two elements. What this thesis proposes is a model that maximizes this correlation, by making accessible to each of the source, cue and content identification steps some information concerning the other two elements. This is achieved by first identifying potential cues and entity-source candidates and using them to derive features to extract the content span.

For the content span extraction, we can already rely on the previously identified cues and on the entities that are recognized by the pre-processing step. These represent potential sources and can be used to derive additional features to drive the content span extraction. By using the cues and the entities as features instead of making a prior hard decision concerning which is the cue and which the entity-source of a given AR, the learning model is allowed more flexibility. It is not bound to a specific cue and content span which could have also been incorrectly identified.

Another advantage of this approach is that it allows building on and comparing to existing literature. Speaker attribution is in fact a well-known task that links a speaker to each given or previously identified quotation. Since quotations are a subset of ARs, similar approaches can be generalized and applied to identify the source span of each attribution content.

4.1.1 Data

Data for training the models had to be built as part of this project, since there were no available large resources annotated with attribution.

An earlier version of the corpus, which is identified as PARC 2.0 (described in Sec. 3.3) and comprises the PDTB ARs I collected and further annotated, was used for the preliminary experiments. This version constitutes a subset of the final corpus: PARC 3.0. A third resource, the SMHC, was also used for developing and testing in part of the experiments. Table 4.1 summarizes the data used in the models described in Pareti et al. (2013) that address the extraction and attribution of ARs of type assertion (PDTB classification), which correspond to quotations. I will discuss PARC 2.0, SMHC and PARC 3.0 here in separate subsections.

PARC 2.0

In the PDTB, ARs were only annotated when they had scope over an entire discourse relation or over one or both of its arguments, leading to a large number of ARs, around half, not being annotated in PARC 2.0. For the initial development of the content extraction and entity attribution components, only non-nested quotations were considered. While incomplete, PARC 2.0 is the only data set that makes the AR type distinction annotated in the PDTB (i.e. assertions, beliefs, facts and eventualities). This was abandoned, due to poor inter-annotator agreement, when completing the annotation of the corpus.

In order to reduce the amount of false negatives coming from unlabelled data in the training set, I looked into different solutions. I found that the verb-cue classifier (described in Sec. 4.1) could be applied to the corpus to identify sentences that were likely to contain an unlabelled attribution. Sentences containing a verb identified as a cue by the classifier and that did not contain a quotation were removed from the training set of the quotation extraction model.

The test set was not affected by this issue, since it corresponds to a fully annotated sample that was double-annotated for the preliminary inter-annotator agreement study (see Sec. 3.2). In this set of 14 articles, both annotators identified 380 ARs, of which 15.5% were nested within another AR. The final test-set includes 267 non-nested ARs of type assertion (i.e. quotations). However, since discontinuous content spans were treated as separate quotations, this led to a slightly larger test-set, totalling 321 non-discontinuous gold quotations (123 direct, 151 indirect and 47 mixed).

SMHC

The second corpus (Pareti et al., 2013)[1] originates from existing annotations of direct quotations within Sydney Morning Herald articles presented in O'Keefe et al. (2012). In that work, quotations were automatically extracted as any text between quotation marks, thus including the directly-quoted portion of mixed quotations, as well as scare quotes. Only quotation speakers were manually annotated. In order to adapt the corpus to the task of extracting all types of quotation spans (i.e. direct, indirect and mixed), one annotator removed scare quotes, completed mixed quotations including both the directly and indirectly quoted portions, and added the indirect quotations. The annotation scheme was developed to be comparable to the scheme used in PARC 2.0 (Pareti, 2012a), although the SMHC corpus only includes quotations (i.e. assertion ARs) and does not annotate the lexical *cue*.

The resulting corpus contains 7,991 quotations taken from 965 articles from the 2009 Sydney Morning Herald and is referred to as SMHC. The annotations in this corpus also include the speakers of the quotations, as well as gold standard Named Entities (NEs). We used 60% of this corpus as training data (4,872 quotations), 10% as development data (759 quotations), and 30% as test data (2,360 quotations). Early experiments were conducted over the development data, while the final results were trained on both the training and development sets and were tested on the unseen test data.

PARC 3.0

The development of PARC 3.0 is described in Ch. 3. The corpus is a superset of

[1] The corpus is presented in joint work. My contribution was to provide the annotation scheme that was used to derive the annotation guidelines. I did not contribute to the corpus collection and annotation.

	SMHC		PARC 2.0	
	Corpus	Doc	Corpus	Doc
Docs	965	-	2,280	-
Tokens	601k	623.3	1,139k	499.9
Quotations	7,991	8.3	10,526	4.6
Direct	4,204	4.4	3,262	1.4
Indirect	2,930	3.0	5,715	2.5
Mixed	857	0.9	1,549	0.6

Table 4.1: Comparison of the SMHC and PARC 2.0 corpora. Document and token size and per quote-type occurrence of quotations for the corpus are reported (Corpus), together with their average per document (Doc).

PARC 2.0 and includes the ARs collected from the PDTB and further annotated as well as the newly annotated ARs (around 50% of ARs were not annotated in the PDTB). The new ARs were annotated on the articles belonging to the news genre[2] which constitutes over 85% of the WSJ corpus.

The fully annotated news section of the corpus was split into training, test and development sets as described in Table 4.2. This follows the standard division adopted by the parsing community (Charniak, 2000) for splitting the WSJ corpus, with the addition of sections 0-1 to the training set. The test section comprises a total of 1,111 ARs. Nested ARs were not part of this study and were therefore excluded from the datasets.

The corpus provided the final data to train, develop and test each component of the model. The model components developed on PARC 2.0 and the SMHC were adapted and extended to address the complete task of extracting not only quotations and their speakers, but all quote types of ARs. The extraction of complete ARs required the addition of new components in order to identify the complete source and cue spans of each AR. The models for the extraction of ARs were developed on the fully annotated corpus: PARC 3.0. The SMHC could not be used as only quotations and speakers are annotated.

[2] A complete list of the WSJ article classified by the genre they belong to can be found here: `http://www.let.rug.nl/~bplank/metadata/genre_files_updated.html`

	TRAIN (0-22)	DEV (24)	TEST (23)	CORPUS
Docs	1706	51	76	1833
Tokens	853k	29k	45k	927k
Verb-cues	18485	630	1127	20242
Mentions	110762	3986	5890	120638
Direct ARs	2936	103	217	3256
Indirect ARs	11045	447	744	12236
Mixed ARs	1978	72	150	2200
Total ARs	15959	622	1111	17692

Table 4.2: Overview of the data from PARC 3.0 used in the experiments. Statistics are shown for the Train, Test and Development sets individually as well as over the whole corpus.

4.1.2 Model Steps

This section presents an overview of the pipeline model summarizing each individual step. The following AR will be used as a running example at each step.

(83) **Jeremiah Mullins, the OTC trading chief at Dean Witter Reynolds in New York**, said proudly *that his company executed every order it received by the close of trading*. (wsj_2379)

1. Preprocessing:

   Data, in the form of documents are pre-processed, adding the required analysis steps using the existing annotation or available tools. In particular, the text is tokenized, lemmatized and POS-tagged, which for PARC 2.0 and PARC 3.0 was done using the available gold standard data from the PDTB while for SMHC by making use of the C&C tools (Curran and Clark, 2003) and the NLTK WordNetLemmatizer Bird et al. (2009). The texts are then parsed using the Stanford Factored Parser (Klein and Manning, 2002) to retrieve both the phrase structure and the dependency trees. NEs are identified using the gold annotations in the BBN corpus (Weischedel and Brunstein, 2005) for PARC 2.0 and PARC 3.0 and the annotated entities in the SMHC. NEs are anonymized in order to prevent overfitting.

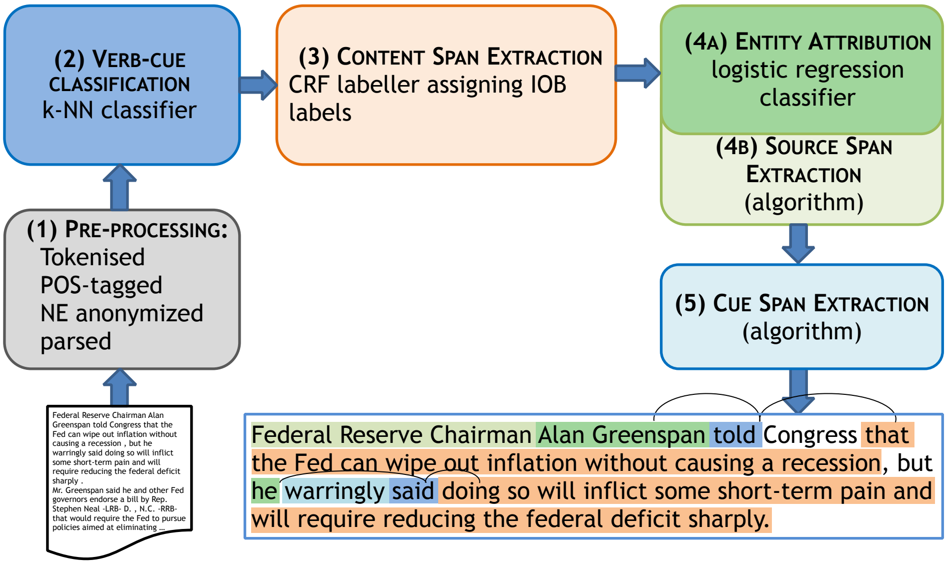

(2) VERB-CUE CLASSIFICATION
k-NN classifier
(3) CONTENT SPAN EXTRACTION
CRF labeller assigning IOB labels
(4A) ENTITY ATTRIBUTION
logistic regression classifier
(4B) SOURCE SPAN EXTRACTION
(algorithm)
(1) PRE-PROCESSING:
Tokenised
POS-tagged
NE anonymized
parsed
(5) CUE SPAN EXTRACTION
(algorithm)
Federal Reserve Chairman Alan Greenspan told Congress that the Fed can wipe out inflation without causing a recession , but he warringly said doing so will inflict some short-term pain and will require reducing the federal deficit sharply .
Mr. Greenspan said he and other Fed governors endorse a bill by Rep. Stephen Neal -LRB- D. , N.C. -RRB- that would require the Fed to pursue policies aimed at eliminating ...
Federal Reserve Chairman Alan Greenspan told Congress that the Fed can wipe out inflation without causing a recession, but he warringly said doing so will inflict some short-term pain and will require reducing the federal deficit sharply.

Figure 4.1: Overview of the model architecture. The constitutive elements of an AR and the model components addressing their extraction are identified by different colours. Cues are light-blue, contents orange and sources green. The source-cue-content triplet that refers to one AR is connected in the example by arches.

2. Verb-cue Classification:

   Head verbs, selected from the gold annotation in the PropBank corpus (Palmer et al., 2005), are classified into attributional, i.e. functioning as the verb-cue of an AR, and non-attributional using the k-nearest neighbour (k-NN) algorithm (Aha and Kibler, 1991). The identified VERB-CUES are then used to derive features for the models in the following steps and as candidate cues for the selection of each AR cue. I developed the classifier on PARC 2.0 and used it to identify cues in PARC 2.0 and in the Sydney Morning Herald Corpus (SMHC) (O'Keefe et al., 2012). I then retrained and applied it to PARC 3.0.

   (84) Jeremiah Mullins, the OTC traiding chief at Dean Witter Reynolds in New York, SAID proudly that his company executed every order it received by the close of trading. (ws_2379)

3. Content Span Extraction:

   The *content* of attribution relations is extracted using a Conditional Random Field (CRF) labeller that assigns *inside* (I), *outside* (O) and *beginning* (B) labels to tokens in a document sequence following the IOB sequence label representation first introduced by Ramshaw and Marcus (1995). This component is part of joint work (Pareti et al., 2013)[3] and was developed on PARC 2.0 and the SMHC. I reapplied the labeller to PARC 3.0 with minor modifications.

(85) Jeremiah Mullins, the OTC traiding chief at Dean Witter Reynolds in New York, SAID proudly *that his company executed every order it received by the close of traiding*. (ws_2379)

4a. Source-Entity Attribution:

Each *content* span is attributed to the **ENTITY** that was assigned the highest probability score by a logistic regression classifier. The original model used was developed on PARC 2.0 and the SMHC by Tim O'Keefe and described in joint work (O'Keefe et al., 2012). I retrained the model on PARC 3.0, after introducing some significant modifications as described in 4.4.3.

(86) **JEREMIAH MULLINS**, the OTC traiding chief at Dean Witter Reynolds in New York, SAID proudly *that his company executed every order it received by the close of traiding*. (ws_2379)

4b. Source Span Extraction:

A set of algorithms is applied to the **ENTITIES** identified by Step 4a to extract the complete AR **source** span. This component was developed on PARC 3.0.

(87) **JEREMIAH MULLINS, the OTC traiding chief at Dean Witter Reynolds in New York**, SAID proudly *that his company executed every order it received by the close of traiding*. (ws_2379)

5. Cue Span Extraction and Linking:

Content and **Source** span pairs are linked to their cue span. The cue span is identified by applying algorithms and using the predictions from Step 2. Cue span modifiers, such as 'proudly' in Ex. (88), are caught as part of the span during the cue span extraction. This component was developed on PARC 3.0. The AR is now complete.

[3] This component, relative to the extraction of quotations, was the result of joint work with Tim O'Keefe. He worked on the implementation of the quotation extraction model and most of the features. My contributions were: (1) the definition of the strategy to learn from the partially unlabelled data of PARC 2.0; (2) the definition of several of the features; (3) the implementation of part of the model and some of the features.

(88) **JEREMIAH MULLINS, the OTC traiding chief at Dean Witter Reynolds in New York**, SAID proudly *that his company executed every order it received by the close of traiding*. (ws_2379)

The model steps are performed in the order presented in this section which is motivated by the reasons discussed at the beginning of Sec. 4.1. Nonetheless, the steps could be arranged in a different order. In particular, the cue span identification, and not the content span, could be the first step since the cue element is the one establishing the relation. The cue span is also more easily identified with respect to the content span, therefore this order could increase the recall for the AR identification. For each cue span we could then identify a source span and at least one content span. However, not every potential cue element will then establish an AR, thus we could expect such order to have a lower precision with respect to the proposed pipeline.

4.1.3 Evaluation Metrics

For the evaluation of the model components, three different metrics were used, as explained in detail below. The first one is a strict metric while the other two account for partially correct predictions. These were used depending on the task. Metrics compare predicted and gold spans. Spans are consecutive sequences of tokens and therefore overlap metrics are not affected by gaps.

- **Strict**

  A span is only considered to be correct if it exactly matches a span from the gold standard. The strict score, however, does not represent well how accurate long-span predictions are. If a prediction is incorrect by as little as one token it will be considered completely incorrect.

- **Partial**

  This is an overlap metric (Hollingsworth and Teufel, 2005), which allows partially correct predictions to be proportionally counted. Taking the sets of gold (gold) and predicted (pred) spans, precision (P), recall (R), and F-score for this method (F) are calculated as follows:

$$P = \frac{\sum_{g \in gold} \sum_{p \in pred} overlap(g,p)}{|pred|} \tag{4.1}$$

$$R = \frac{\sum_{g \in gold} \sum_{p \in pred} overlap(p,g)}{|gold|} \tag{4.2}$$

$$F = \frac{2PR}{(P+R)} \tag{4.3}$$

$overlap(x,y)$ returns the proportion of tokens of y that are overlapped by x. For each of these metrics micro-average scores are reported, as the number of ARs in each document varies significantly. When reporting results on the different AR quote types, we restrict the set of predicted and gold ARs to only those with the requisite quote type.

- **Soft**

  The third is also an overlap metric, which takes into account partial matches by considering correct a prediction having any overlap with the gold span. While this is not a good metric for longer spans, such as content spans, it is a good indicator for short sequences of tokens. The Soft metric was used to evaluate the cue span prediction, where spans are usually one or two tokens long. If the algorithm fails to recognize a modifier, e.g. 'repeatedly' in 'repeatedly said', the partial metric would heavily penalize this, while the soft metric would count this as correct.

The metrics were used to calculate *precision*, *recall* and *F-score*. When this was not meaningful, since the model would make a prediction for each gold span, as in the case of source and cue span identification, *accuracy* was calculated.

The *quote status* of an attribution, i.e. whether direct, indirect or mixed, determines the different structures that can carry it and therefore its complexity for the identification and extraction task. While the content span of direct AR can more often span over sentences, it is enclosed by quotation marks and therefore relatively trivial to identify and extract with punctuation clues. An indirect AR instead is much harder to identify and its content span boundaries more complex to determine with precision because of ambiguities in the underlying syntactic structure.

In order to evaluate the models on the different AR quote statuses, part of the results will be presented for *direct* (D), *indirect* (I) and *mixed* (M) separately. This will enable quantifying the intrinsic complexity of each quote status and compare results on the different corpora that have a different proportion of ARs per quote status.

The quote status of each AR, whether gold or predicted, is automatically calculated using Algorithm 2 by looking at the presence of quotation marks in the content span.

Since incorrect predictions can lead to a predicted content span being identified as having a different quote status than the gold span it matches, when a predicted AR matches a gold AR it inherits its quote status. The calculated quote status for the predicted ARs is used to add the false positives to the results of the respective quote status.

Algorithm 2 Attribution quote status assignment

```
procedure SETARQUOTE STATUS($span$)    ▷ Set the quote status of an AR given its content span
    if $span.startToken = quotMark$ and $span.endToken = quotMark$ then
        $quotestatus \leftarrow direct$
    else if any token in span = quotMark then
        $quotestatus \leftarrow mixed$
    else
        $quotestatus \leftarrow indirect$
```

Statistical Significance

I run statistical tests on the models of each attribution extraction component in order to determine whether the difference between models and baselines was statistically significant. Tests on the SMHC could not be run since the original data was not available.

For the binary predictions, significance is calculated using McNemar's Chi-square (Binomial Test) test[4] for paired categorical data (McNemar, 1947). This applies to the verb-cue classifier as well as to the strict and soft metrics of all other models. For the partial metrics, I used the non-parametric Wilcoxon signed-rank test (Wilcoxon, 1945)[5] for non nominal data.

While the verb-cue classification and source-entity attribution models make predictions on a fixed set of items, the content span extraction model makes free predictions, namely the set of spans is not a pre-defined list. This model is also evaluated with respect to the gold spans for that task, however, gold spans are only positive instances since there is not a fixed set of positive and negative spans. In order to take false predictions into account, predictions not matching a gold span were added to the gold list as

[4] Implementation by Ernesto P. Adorio: mcnemar.py. Available at: http://adorio-research.org/wordpress/?p=238 (Accessed 5 February 2015).

[5] SciPy (Jones et al., 2001) implementation: scipy.stats.wilcoxon. Available at: http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.wilcoxon.html (Accessed 5 February 2015).

Instance	Baseline	Model
True Gold 1	0	1
True Gold 2	1	1
True Gold 3	0	1
False Pred 1	1	0
False Pred 2	0	0

Table 4.3: Statistical Significance Scoring Example.

negative instances and each model scored accordingly, i.e. 0 if it predicted the negative instance, 1 if it did not. As shown in Table 4.3, a false prediction made by the model (False Pred 1) is added to the list of instances as a false instance and the Baseline is rewarded for not predicting it. When both models make the incorrect prediction, they are both assigned a score of 0 (False Pred 2).

4.2 Cue Identification

Attribution cues carry information that is fundamental for a correct understanding of the AR and contribute to determining its features: type (e.g. 'declare' vs. 'think'); factuality (e.g. 'say' vs. 'didn't say'); scopal change (e.g. 'declare' vs. 'deny'); authorial stance (e.g. 'claim' vs. 'confirm') and source attitude (e.g. 'smile' vs. 'purr'). In addition to the element expressing the cue, usually a verb, the cue span includes particles, modal verbs and modifiers such as 'said grudgingly' or 'would think', that contribute to the interpretation of the AR.

Being the lexical anchor of the AR, cues can strongly contribute to the identification of its content and source spans, which are often syntactically and semantically related to the cue. Despite being crucial to AR extraction, the literature has so far underestimated their importance. Cues are commonly used to detect the quotation and the speaker, however quotation attribution studies only establish a link between the quotation and the speaker without retaining the cue. The nature of the relation which the cue expresses is therefore lost. What this means is that we are not able to distinguish whether the source 'said' or 'didn't say', 'confessed' or 'denied' the attributed quotation.

The identification of cues is the first step towards building a pipeline model for labelling ARs. This Section addresses the identification of potential cue elements. These cues are then used in subsequent steps as features to guide the identification of source and content spans. Once source and content of an AR have been extracted, the cue for that AR is selected from the identified cue candidates. The selection of the cue element and the extraction of the complete cue span, including its modifiers are tackled as a separate step in Sec. 4.5.1.

4.2.1 Related Work: Relying on Lists of Reporting Verbs

The verb-cue has a central role for the identification of an attribution content and source, whether as a feature of a ML system or as part of a rule-based pattern. As a consequence, several attribution studies have devised a strategy to recognise these verbs (see Sec. 2.2 for a full discussion of the literature).

Mostly, this consisted of a small expert-derived list covering only the most frequent reporting or opinion verbs:

- 27 verb types in Bethard et al. (2004);
- 35 in Sarmento and Nunes (2009);
- 54 in Krestel et al. (2008) (these are reported in Table 4.4);
- 114 in de La Clergerie et al. (2009).

More systematic semantically-driven approaches where adopted by Lu (2010); Das and Bandyopadhyay (2010); Elson and McKeown (2010). Lu (2010) collected 308 verbs by searching available lexical resources for synonyms of the original 68 verbs collected from their dataset.

Using WordNet, Das and Bandyopadhyay (2010) and Elson and McKeown (2010) compiled a list of affect and expression verbs respectively. In the latter, this includes over 6,000 tokens (which comprise capitalised and conjugated verb forms).

Lists have however two major drawbacks:

- Incomplete: attributional verbs follow a sort of Pareto principle, the 80-20 rule, or Zipfian distribution, with the first 10% of types covering around 90% of the occurrences. This means that even fairly short lists can provide broad coverage. However, this top 10% is highly dependent on genre, domain and stylistic differences (see Sec. 5.2.1) and there is a long tail of less common or infrequently occurring verbs that cannot be exhaustively listed.

according	accuse	acknowledge	add	admit
agree	allege	announce	argue	assert
believe	blame	charge	cite	claim
complain	concede	conclude	confirm	contend
criticize	declare	decline	deny	describe
disagree	disclose	estimate	explain	fear
hope	insist	maintain	mention	note
order	predict	promise	recall	recommend
reply	report	say	state	stress
suggest	tell	testify	think	urge
warn	worry	write	observe	

Table 4.4: Set of reporting verbs defined by Krestel et al. (2008)

- Imprecise: taking every occurrence of a verb as attributional has a major negative impact on precision. Even highly predictive attributional verbs, e.g. 'say' can be used in a non attributional context (e.g. 'Well said!', 'That said, ...'). Moreover, some of the most frequent attributional verbs are very common verbs that only occasionally pair up with attribution (e.g. 'add', 'continue' and 'show'). Contextual information is therefore necessary to disambiguate among different senses and uses of each verb.

4.2.2 Verb-cue Classification

Verbs are by far the most common introducer of an AR. In PARC 3.0, verbs account for 92% of all cues, the prepositional phrase *according to* for 2%, with the remaining 6% being nouns, adverbials and prepositional groups (see Sec. 3.5.2.1). Attributional verbs are not a closed set and their occurrence and frequency vary depending on: genre, domain, register and style (see Sec. 5.2).

It is therefore not possible to simply rely on a pre-compiled list of common speech verbs. Quotations in PARC 3.0 are introduced by 527 verb types, 199 of which are unique occurrences. Not all of the verbs are speech verbs as a range of non-reporting verbs can in some contexts or occasionally have an attributional use, for example *add* (Ex. (89)), which is the second most frequently attributional verb after *say*, or verbs such as *gripe*, *smile* and *fume*.

(89) a. In his ruling, Judge Curry added an additional $ 55 million to the commission's calculations. (wsj_0015)

b. The bids, he added, were "contrary to common sense." (wsj_0051)

This section describes a supervised classifier that can automatically identify cue-verbs, i.e. whether each occurrence of a verb is used as attributional.

4.2.2.1 Model

The attributional cues annotated in PARC 2.0 were used to develop a separate component of the system that identifies attribution verb-cues. The system was then retrained on PARC 3.0, using all files and ARs in the training portion of the corpus.

The binary classifier predicts whether the head of each verb group is a verb-cue. Verb group heads were identified by selecting the verbs annotated in the PropBank corpus (Palmer et al., 2005). The classifier consists in the Weka (Hall et al., 2009) implementation of the k-nearest neighbour (k-NN) algorithm. The algorithm labels each test instance, represented as an unlabelled vector in a multidimensional feature space, by looking at the *k* closest training examples, represented as labelled vectors. The class label is assigned to the instance by majority vote. The optimal number of neighbour instances for this task was equal to 4 (ties were classified as non-attributional). The algorithm uses Euclidean distance and no distance weighting as this proved to slightly improve recall, but had a strong negative impact on precision.

Table 4.5 shows the proportion of verb-cues over head verbs in PARC 3.0. This proportion is extremely high, with over 23% of head verbs in the Test set and 21% in the Dev set being cues of an AR. This proportion is lower on the training set simply because only around half of this section of the corpus has been annotated with nested ARs. The training set was sampled to 35% of its original size, in order to have a more uniform distribution of positive and negative instances, i.e. positive instances were retained while negative instances were reduced to a similar number by randomly sampling them. The result is a more balanced training set with 45% of positive instances instead of just 16% of the original training set (Table 4.5).

4.2.2.2 Features

Features are a combination of lexical, syntactic and semantic features that contribute to defining the context of a verb and whether it can assume an attributional meaning.

SECTION	HEAD VERBS	VERB-CUES	VERB-CUES RATE
Train (0-22)	102,617	16,351	16%*
Train sampled	35,915	16,351	45%**
Dev (24)	3,265	685	21%
Test (23)	5,282	1,233	23%

Table 4.5: Verb-cues per corpus section. (*)The rate of verb-cues for the training section is lower due to nested attributions being annotated in only half of the documents.(**) The training set was sampled to have a more uniform distribution of verb-cues over non-cue verbs.

This is achieved by identifying the VerbNet (Kipper et al., 2000) classes a verb is a member of. Verbnet is a lexicon of verbs organised in classes and subclasses. These are an extension of Levin (1993) and semantically group verbs. For each head of verb group, the classifier uses 20 feature types:

- Lexical: token, lemma, next/previous token.
- Punctuation: colon/quotation mark adjacency.
- Grammatical: POS of next/previous token.
- VerbNet classes membership: binary feature identifying each VerbNet class a verb is part of (e.g. 'support' is a member of admire-31.2, contiguous_location-47.8 and help-72).
- Syntactic: node-depth in the sentence, parent node and parent sibling nodes.
- Sentence features: distance from sentence start/end, within quotation marks.

4.2.2.3 Baselines

The classifier is compared against two baselines. The first one, B*say*, marks every occurrence of *say* as positive, i.e. verb-cue. *Say* is overwhelmingly predominant as a verb-cue and considered as very accurate. The second baseline, B*list*, marks as positive every occurrence of each of the 54 verbs (Table 4.4) from the expert-compiled list in Krestel et al. (2008) which was used to extract reported speech from the WSJ corpus. This baseline allows to evaluate how effective manually collected lists are.

4.2.2.4 Results

The results in Table 4.6 show that the verb-cue classifier outperforms expert-derived knowledge. The classifier was able to identify verb-cues with 84% *precision* and 86% *recall*, i.e. an improvement of 10% *precision* and 16% *recall* with respect to the list of verbs used in Krestel et al. (2008). While the results show that frequently occurring verbs cover the bulk of the instances and are relatively highly predictive, they also show how a supervised model with the inclusion of VerbNet classes (Schuler, 2005) and contextual features is both more precise and achieves better coverage.

Such a model, in particular, allows for a more accurate classification of polysemous verbs. For example, of the 38 occurrences of 'add' in the test set, 60% are cues and 40% correspond to other uses. While the baseline (B*list*) will label all instances as cues, the classifier correctly recognizes all attributional uses and misclassifies only two non-attributional instances. Similarly, the 16 occurrences of 'feel' are correctly identified as attributional in 50% of the cases, with one misclassified negative instance.

The classifier is also able to correctly label some rarely occurring or unseen verbs, such as the ones in Ex. (90), which are not on the list in Table 4.4.

(90) a. Today, he frets, exports and business investment spending may be insufficient to pick up the slack if stock prices sink this week and if consumers retrench in reaction. (wsj_2397)

b. Many of the nation's highest-ranking executives saluted Friday's market plunge as an overdue comeuppance for speculators and takeover players. (wsj_2345)

Although very simple, the model is also able to recognize some of the non attributional uses of 'say', in spite of it being almost exclusively (around 97% of the times) used as attributional. For example, the occurrence of 'say' in Ex. (91) is correctly labelled as not being a cue.

(91) ... if Mr. Mason's type of ethnic humor is passe, then what other means do we have for letting off steam?

Don't say the TV sitcom, because that happens to be a genre that, in its desperate need to attract everybody and offend nobody, resembles politics more than it does comedy. (wsj_2369)

	Precision	Recall	F-score	Accuracy
B*say*	97	45	61	87
B*list*	74	70	72	87
k-NN	84	86	85	93

Table 4.6: Comparison of the results for the verb-cue classification task. All differences between models are statistically significant, with two-tailed P value less than 0.0001 (McNemar's test).

The developed verb-cue classifier was then applied to the SMHC in order to identify the verb-cues in that corpus, since the SMHC does not annotate AR cues. Head verbs were identified by taking the verb element of each verb phrase having no other verb phrase as direct child. This constraint excludes auxiliaries and modal verbs (e.g. *(VP (VBP have)(VP (VBD said)))*. The identified verb-cues were used in subsequent models to derive features or algorithms.

Algorithm 3 Head of Verb Phrase Identification

```
procedure GETHEADVERBS(document)            ▷ identify all verb phrase (VP) heads
    for VP in document do
        if not VP has another VP as direct child then
            for all children of VP do
                if child is terminal node and child.PoS starts with VB then
                    add child to head verbs
```

4.2.3 Recognising Other Cues

Verbs are by far the most common type of attributional cues, however, the cue element might not be a verb. Other type of cues are heterogeneous and relatively infrequent, which hinders the development of supervised models. These cues might be expressed by punctuation, nouns, adjectives and prepositional groups, through a range of different structures.

In order to partially overcome the limitation of only relying on verbal cues, I have produced a list of attributional nouns which is derived from the attributional verbs (e.g. 'report', 'thought', 'intention') and integrated with additional potentially reporting nouns (e.g. 'letter', 'words', 'idea'). The complete list is reported in Appendix

C. With respect to the PARC 3.0 corpus, the list covers 69% of noun cue occurrences, while most non de-verbal nouns or less frequent nouns are not identified (e.g. 'admonition', 'principle', 'unwillingness').

In addition to nouns, occurrences of 'according to' are also classified as attributional, since it is the most productive cue apart from verbs and it is extremely precise. Nonetheless, in rare cases, 'according to' has a non attributional use as in Ex. (92a), which differs only semantically from the attributional use (Ex. (92b)).

(92) a. They painted the apartment orange, pink and white, according to her instructions. (wsj_2343)

b. *They painted the apartment orange, pink and white,* according to **her letter**.

4.3 Content Extraction

4.3.1 Assertion Attributions[6]

Attribution is not a homogeneous field and different aspects of it have been the object of separate groups of studies. Particular interest and effort have been directed to the extraction of one type of attribution, namely quotations or reported speech (see Sec. 2.2.1). Based on the type of attitude the source expresses towards a proposition or eventuality, attributions in the PDTB are subcategorised (Prasad et al., 2006) into *assertions* (Ex. 93a) and *beliefs* (Ex. 93b), which imply different degrees of commitment to the truth of the proposition, *facts* (Ex. 93c), expressing evaluation or knowledge, and *eventualities* (Ex. 93d), expressing intention or attitude.

(93) a. Mr Abbott said *that Arnold is a lawyer.*

b. Mr Abbott thinks *that Arnord is a lawyer.*

c. Mr Abbott knew *that Gillard was in Sydney.*

d. Mr Abbott agreed *to the public sector cuts.*

[6] (Joint work)Parts of this section are based on work published in Pareti et al. (2013). This was joint work with Tim O'Keefe, who provided the SMHC corpus, worked on the implementation of the quotation extraction model and most of the features and developed the source-attribution component of the system. I provided the PARC 2.0 corpus, developed the verb-cue classifier component and the strategy to learn from PARC 2.0 partially unlabelled data and contributed to the definition and implementation of the approach and the features.

	Method	LC	Test Size	Results	
			(quotations)	*P*	*R*
Krestel et al. (2008)	hand-built grammar	EN	133	74	99
Sarmento and Nunes (2009)	patterns over text	PT	570	88	5*
Fernandes et al. (2011)	ML and regex	PT	205	64†	67†
de La Clergerie et al. (2009)	patterns over parse	FR	40	87	70
Schneider et al. (2010)	hand-built grammar	EN	N/D	56†	52†

Table 4.7: Related work on direct, indirect and mixed quotation extraction. Note that the results are not directly comparable as they apply to different languages and greatly differ in evaluation style and size of test set. Language code (LC): English (EN), French (FR), Portuguese (PT). *Figure estimated by the authors for extracting 570 quotations from 26k articles. †Results are for quotation extraction and attribution jointly.

Only assertion attributions necessarily imply a speech act. Their *content* corresponds to a quotation span and their *source* is generally referred to in the literature as the *speaker* and sometimes as the *author* (see Sec. 1.3).

Assertion attributions are the most common type of attribution. They are easier to recognize since a fair proportion of them is represented by direct quotations, whose content span is delimited by quotation marks. Quotation attribution is a well-attested field in the literature. In order to compare to these studies and draw on a common ground, we have first addressed the extraction of the content of assertions only and afterwards extended our models to cover all attribution types.

Direct, indirect and mixed quotations differ in the degree of factuality they entail, since direct quotations are by convention interpreted as a verbatim transcription of an utterance whereas indirect and the non-quoted portion of mixed quotations can be paraphrased forms of the original wording, and are thus more likely to have been modified by the writer. Direct quotation attribution, with direct quotations being given or extracted heuristically, has been the focus of studies in both the narrative (Elson and McKeown, 2010) and the news (Pouliquen et al., 2007; Liang et al., 2010) domains (see Sec. 2.2). The few studies that have also addressed the extraction and attribution of indirect and mixed quotations are summarized in Table 4.7.

This shows that the majority of evaluations so far have been on a small-scale. Furthermore, the published results do not include any comparisons with previous work, which prevents a quantitative comparison of the approaches, and they do not include results broken down by whether the quotation is direct, indirect, or mixed. This is particularly relevant due to the inherent lower complexity of detecting direct quotations, whose proportion in the test data can have a great impact on the overall results.

Fernandes et al. (2011) is the closest to the proposed approach as they partially apply supervised machine learning to quotation extraction. They treat quotation extraction as an IOB labelling task, where they use the Entropy Guided Transformation Learning (ETL) algorithm (Santos and Milidiú, 2009) with POS and NE features to identify the beginning of a quotation, while the inside and outside labels are found using regular expressions. Finally they use ETL to attribute quotations to their source. The overall system achieves 64% *precision* and 67% *recall*.

The token-based approach (**Token**) treats quotation extraction as analogous to sequence tagging, where there is a sequence of tokens that need to be individually labelled. Each token is given either a B, an I, or an O label, where B denotes the first token in a quotation, I denotes the token is inside a quotation, and O indicates that the token is not part of a quotation.

For NE tagging it is common to use a sentence as a single sequence, as NEs do not cross sentence boundaries. This does not work for quotations, as they can cross sentence and even paragraph boundaries. As such, we treat the entire document as a single sequence, which allows the predicted quotations to span both sentence and paragraph boundaries.

As the learning algorithm, we use Okazaki (2007) implementation of linear chain Conditional Random Field (CRF) (Lafferty et al., 2001).

4.3.1.1 Features

The features used for the Token model are a combination of lexical, syntactic and positional features. We selected a broad range of features that could help the identification of the source span boundaries. Lexical features encode lexical and grammatical information of the tokens within a certain window from the target one as well relative to the whole sentence including it. These features can encode the presence of quotation marks and cues as well as potential entities and may be already rather effective for direct quotations.

Since indirect quotations are more strongly connected with the syntactic level and tend to respect constituent boundaries, we also added dependency and syntactic features to the model. These features include the position of the target token within a constituent since this could correspond to the beginning of a quotation. We included dependency features to account specifically for the relation of the target with a potential verb-cue.

The interconnection of source, cue and content suggested the inclusion of features related also to the source. Knowledge of the presence of a NE or a pronoun was integrated with external knowledge to encode the presence of titles, roles and organisation names. Titles come from a small hand-built list. Lists of roles and organisations were built by recursively following the WordNet (Fellbaum, 1998) hyponyms of person and organization respectively. The features we implemented for the Token model are summarised below:

Lexical: unigram and bigram versions of the token, lemma and POS tags within a window of 5 tokens either side of the target, all indexed by position.

Sentence: features indicating the sentence length and whether it contains a quotation mark, a NE, a verb-cue, a pronoun, or any combination of these.

Dependency: relation with parent, relations with any dependants, as well as versions of these that include the head and dependent tokens.

Verb: features indicating whether the current token is a (possibly indirect) dependent of a verb-cue, and another for whether the token is at the start of a constituent that is a dependent of a verb-cue.

Syntactic: the label, depth, and token span size of the highest constituent where the current token is the left-most token in the constituent, as well as its parent, and whether either of those contains a verb-cue. The labels of all constituents that contain the current token in their span, indexed by their depth in the parse tree.

External knowledge: position-indexed features for whether any of the tokens in the sentence match a known role, organisation, or title.

Other: features for whether the target is within quotation marks and whether there is a verb-cue near the end of the sentence.

4.3.1.2 Baselines

We have developed three baselines inspired by the current lexical/syntactic pattern-based approaches in the literature, which combine speech verbs and hand-crafted rules. Although these approaches are very simple, they provide a comparison term to evaluate our methods against previous studies. They allow us to measure, on a large scale and over different corpora, the real predictive power of such rules and the actual gain of a machine-learning approach.

Punctuation (B_{pun}) The first baseline is based on punctuation and addresses the identification of direct quotations only. Direct quotations have been the starting point of many speaker attribution studies as they are the least challenging to identify. Punctuation clearly draws the boundaries of the content span which is enclosed by quotation marks as in Ex. (94). Depending on the convention adopted, for quotations spanning over paragraphs an opening quotation mark can be repeated at the beginning of each paragraph.

(94) “I just don’t feel that the company can really stand or would want a prolonged walkout,” Tom Baker, president of Machinists’ District 751, said in an interview yesterday. “I don’t think their customers would like it very much.” (wsj_2308)

Although a seemingly trivial task, a few challenges had to be addressed:

1. Scare Quotes: These surround words or phrases to imply a different reading than the commonly associated one (e.g. ‘They established a “non-profit” organization’).

   Solution: We set a length requirement and discard quoted spans shorter than three tokens. Although not very frequent, short quotations can occur and are erroneously discarded by this rule while longer scare quotes are also mistakenly identified. Nonetheless, we identified this length limit as the best trade-off. For PARC 3.0, this length limit was raised to 5, as this proved to yield the best results. Fig. 4.2 shows the results for different tested values (between 1 and 8) of the minimum required span length. The results refer to direct ARs only. The quoted portion of mixed ARs that the baseline erroneously identifies as direct is not considered in these results in order to decouple their detrimental effect from that of scare quotes.

2. Titles: Book and film titles are also identified by quotation marks (e.g. ‘She read “From Here to Eternity”’)

   Solution: Discarding quoted spans having all non-mention and non sentence-initial words capitalized, with the exception of stopwords. When there were no such words in the span, the quotation was retained.

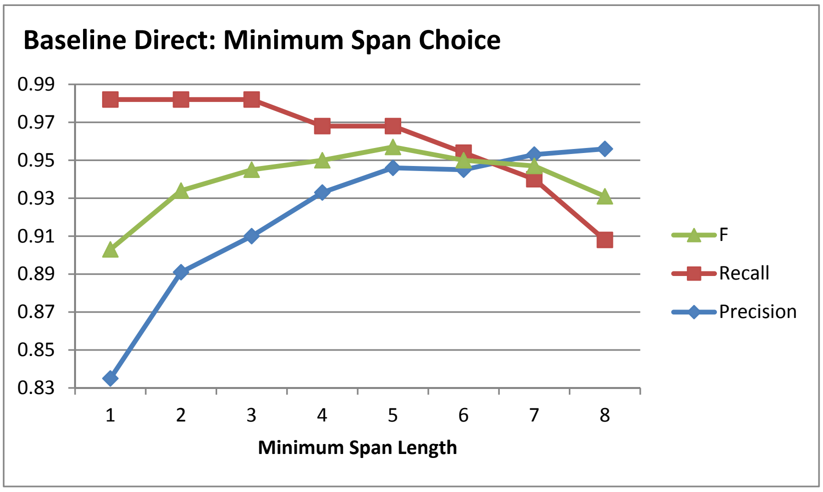


Figure 4.2: Comparison on PARC 3.0 of the effect of choosing a different minimum length value for the quoted span extracted by the Punctuation (B_{pun}) baseline. For PARC 3.0, the optimal minimum length is 5, as this is the best trade-off between precision and recall. The quoted portion of mixed ARs that the baseline identifies are not included in the results in order to evaluate the effect of scare quotes only.

3. Quoted span within a quotation: Conventionally quotation marks alternate between double (“ ”) and single (‘ ’) to indicate the presence of a quotation within a quotation such as: “The man was drunk and shouted: ‘You have to follow your dreams’” he told the police. In addition, different quoting conventions might need to be taken into account.

   Solution: Since we did not include nested quotations in the scope of our model, this could be disregarded. Quoting style was consistent throughout our data and only (“ ”) and (‘ ’) were found.

Lexical (B_{lex}) In order to identify also indirect quotations (Ex. (95)) and the non quoted portion of mixed quotations (Ex. (96)), we adopted a baseline that extracts the longest of the spans between a verb-cue and either of the sentence boundary. Although void of linguistic knowledge, this baseline can already identify the content of a large portion of ARs that span over a complete sentence. In the following examples, it would correctly identify the content of Ex. (95) and (96) but not the content of Ex. (97), since the span following the verb-cue is one token longer than the span preceding it.

(95) He added *that the company miscalculated the union’s resolve and the workers’*

B_{pun}	Punctuation: span within quotes, longer than 3 tokens, excluding titles.
B_{lex}	Lexical: longest span between cue-verb and sentence boundary.
B_{syn}	Syntactic: syntactic object of the cue-verb.

Table 4.8: Baselines for the content extraction task

disgust with being forced to work many hours overtime. (wsj_2308)

(96) A Boeing spokeswoman said *a delivery date for the planes is still being worked out "for a variety of reasons, but not because of the strike."* (wsj_2308)

(97) "We want to make sure they know what they want before they come back," said *Doug Hammond, the federal mediator who has been in contact with both sides since the strike began.* (wsj_0472)

Syntactic (B_{syn}) The third baseline takes the syntactic object of the verb-cue as the content span. This baseline is inspired by current syntactic rule-based Attribution extraction systems, e.g. de La Clergerie et al. (2009). These approaches start from a given verb and identify its source and content spans by retrieving the verb's syntactic subject and object respectively. We identify the syntactic object using Stanford dependencies by taking the clausal complement (ccomp) of the cue-verb. Although *ccomp* is not the only relation type that could establish verb-object relationship, the addition of more types (e.g. dobj) proved detrimental.

Instead of relying on a lexicon of verbs, our baselines use those identified by the verb-cue classifier.

4.3.1.3 Results

Direct Quotations

Table 4.9 shows the results for predicting direct quotations on PARC 2.0 and SMHC. In both corpora and with both metrics the token-based approach outperforms B_{pun}. Although direct quotations should be trivial to extract, and a simple system that returns the content between quotation marks should be hard to beat, there are two main factors that confound the rule-based system.

The first is the presence of mixed quotations, which is most clearly demonstrated in the difference between the strict precision scores and the partial precision scores for

		Strict			Partial		
		P	R	F	P	R	F
PARC 2.0	B_{pun}	75^{+}	94	83^{+}	96^{+}	94	95^{+}
	Token	$97^{\bullet}$	91	$94^{\bullet}$	$98^{\bullet}$	97	$97^{\bullet}$
SMHC	B_{pun}	87	93	90	98	94	96
	Token	94	90	92	99	97	98

Table 4.9: PARC 2.0 and SMHC results on direct quotation extraction. The token based approach is trained and tested on all quotations. ($^{\bullet}$: significantly different from B_{pun}; $^{+}$: significantly different from Token).

B_{pun}. B_{pun} will find all of the directly-quoted portions of mixed quotations, which do not exactly match a quotation, and so will receive a low precision score with the strict metric. However, the partial overlap score will reward these predictions, as they do partially match a quote, so there is a large difference in those scores. Note that the reduced strict score does not occur for the token method, which correctly identifies mixed quotations. Mixed quotations are a much higher proportion in PARC 2.0 (1:2) than in the SMHC (1:5), which explains the much lower strict precision in PARC 2.0.

The other main issue is the presence of quotation marks around items such as book titles and scare quotes. In Section 4.3.1.2 we described the methods that we use to avoid scare quotes and titles, which are rule-based and imperfect. While these methods increase the overall F-score of B_{pun}, they do have a negative impact on recall. For PARC 2.0 this could be quantified as a 7% drop in recall while the gain in precision was of 14%. These results demonstrate that although direct quotations can be accurately extracted with rules, the accuracy will be lower than it might be anticipated and the returned spans will include a number of mixed quotations, which will be missing some content.

Indirect and Mixed Quotations

The token approach was also the most effective method for extracting indirect and mixed quotations as Tables 4.10 and 4.11 show. Indirect quotations were extracted with strict F-scores of 59% and 60% and partial F-scores of 76% and 74% in PARC 2.0 and SMHC respectively, while mixed quotations were found with strict F-scores of 56% and 85% and partial F-scores of 87% and 86%.

Although there is a strong interconnection between syntax and attribution, results for B_{syn} show that merely considering attribution as a syntactic relation (Skadhauge

	Indirect			Mixed			All		
Strict	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
B_{lex}	$34^{+\bullet}$	$32^{+\bullet}$	$33^{+\bullet}$	$17^{+\bullet}$	26	$20^{+\bullet}$	$46^{+\bullet}$	$44^{+\bullet}$	$45^{+\bullet}$
B_{syn}	$78^{\diamond}$	$46^{\diamond}$	$58^{\diamond}$	$61^{\diamond}$	40	$49^{\diamond}$	$80^{\diamond}$	$63^{\diamond}$	$70^{\diamond}$
Token	$66^{\diamond}$	$54^{\diamond}$	$59^{\diamond}$	$55^{\diamond}$	58	$56^{\diamond}$	$76^{\diamond}$	$70^{\diamond}$	$73^{\diamond}$
Partial	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
B_{lex}	$56^{+\bullet}$	66^{+}	$61^{+\bullet}$	$78^{+\bullet}$	79	$78^{+\bullet}$	$73^{+\bullet}$	79	$76^{+\bullet}$
B_{syn}	$89^{\diamond}$	58^{+}	$70^{\diamond}$	$88^{\diamond}$	75	$81^{\diamond}$	$92^{\diamond}$	74	$82^{\diamond}$
Token	$79^{\diamond}$	$74^{\diamond\bullet}$	$76^{\diamond}$	$85^{\diamond}$	90	$87^{\diamond}$	$87^{\diamond}$	86	$87^{\diamond}$

Table 4.10: PARC 2.0 results on quotation extraction. *All* reports the results over all quotations (direct, indirect and mixed). For the baselines, this is a combination of the strategy in B_{lex} or B_{syn} with the rules for direct quotations. ($^{\bullet}$: significantly different from B_{syn}; $^{\diamond}$: significantly different from B_{lex}; $^{+}$: significantly different from Token).

	Indirect			Mixed			All		
Strict	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
B_{lex}	37	42	40	15	36	21	50	50	50
B_{syn}	63	49	55	67	36	47	82	72	76
Token	69	53	60	80	91	85	82	75	78
Partial	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
B_{lex}	52	68	59	87	77	82	77	84	81
B_{syn}	75	59	66	89	66	76	91	80	85
Token	82	67	74	88	84	86	92	86	89

Table 4.11: SMHC results on quotation extraction. *All* reports the results over all quotations (direct, indirect and mixed). For the baselines, this is a combination of the strategy in B_{lex} or B_{syn} with the rules for direct quotations.

and Hardt, 2005) has a large negative impact on recall because only a subset of inter-sentential quotations can be effectively matched by verb complement boundaries. B_{syn} is particularly effective on indirect quotations. In PARC 2.0 it yields similar results to the Token model for the strict score, with lower recall (46% vs. 54%) but higher precision (78% vs. 66%). While the recall suffers from considering only verb-cues, this increases precision since quotations having a verb-cue tend to consistently express the content as a clausal complement.

For mixed quotations, B_{syn} scores are considerably lower than the model. What this suggests is that indirect quotations are more syntactically encoded than mixed ones. This is not surprising since mixed quotations, similar to direct ones, can make use of quotation marks to enclose part of the content span. This can span over sentences as in Ex. (98) and thus not strictly match syntactic verb arguments.

(98) *In fact, "the market has always tanked. Always. There's never been an exception,"* says **Gerald W. Perritt, a Chicago investment adviser and money manager**, based on a review of six decades of stock-market data. (wsj_0090)

In the SMHC, B_{syn} is instead considerably worse than the model, including precision scores. We have to consider that the verb-cues were extracted using the verb-cue classifier developed on PARC 2.0 and that the corpus annotation may be less thorough. Thus the baseline might identify verbs that are not associated with attribution in the corpus due to stylistic differences as well as quotations that have not been annotated.

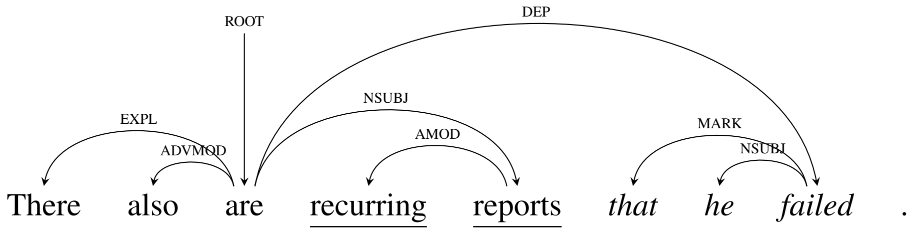
Tables 4.10 and 4.11 also report results for the extraction of all quotations, irrespective of their type. For this score, the baseline models for indirect and mixed quotations are combined with B_{pun} for direct quotations.

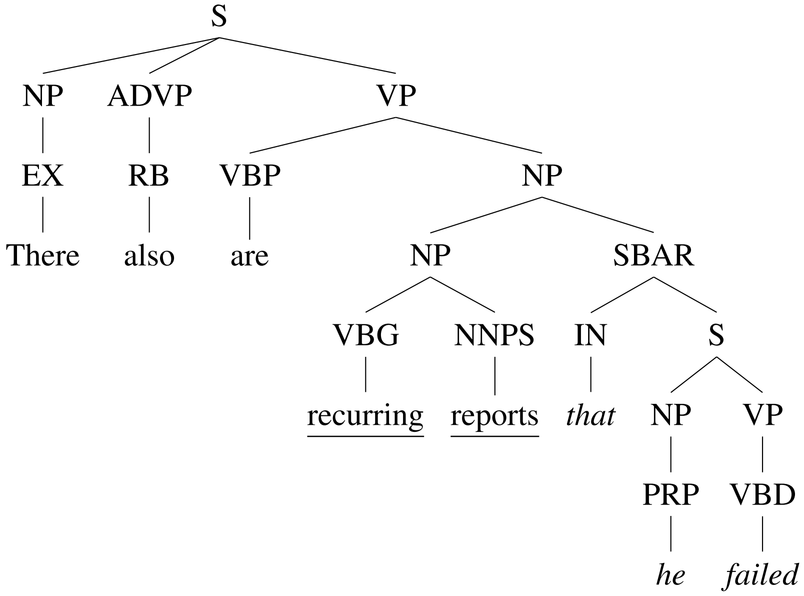
4.3.2 All Types of Attribution

In order to extract also other types of ARs, I adapted and applied the token-based approach to the full corpus (PARC 3.0), which comprises not only assertion ARs, but also beliefs, facts and eventualities as in the original PDTB annotation. The identification of nested ARs was considered as a separate task and therefore nested ARs were ignored (see Sec. 6.2 for a discussion on extracting nested ARs).

In order to address some of the sources of errors and limitations of the token-based model developed for quotations, the following modifications were introduced:

- Addition of sentence, node and syntactic features for the manually collected attributional nouns (Sec. 4.2.3) in order to capture the relation between the token and these potential cues which are not easily captured by dependency features. The example below shows the dependency and syntactic structure for an AR with a noun cue: ‘recurring reports’. The clausal content span is not in a dependency relation with the cue while it is syntactically part of the same noun phrase.

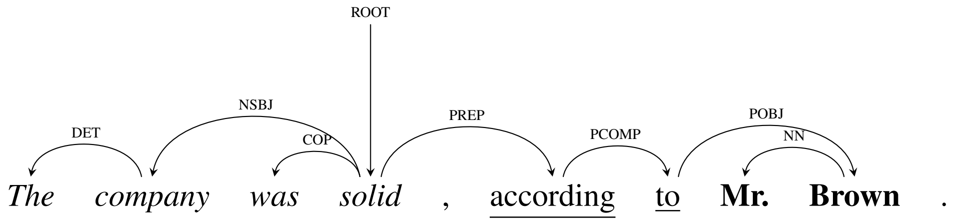


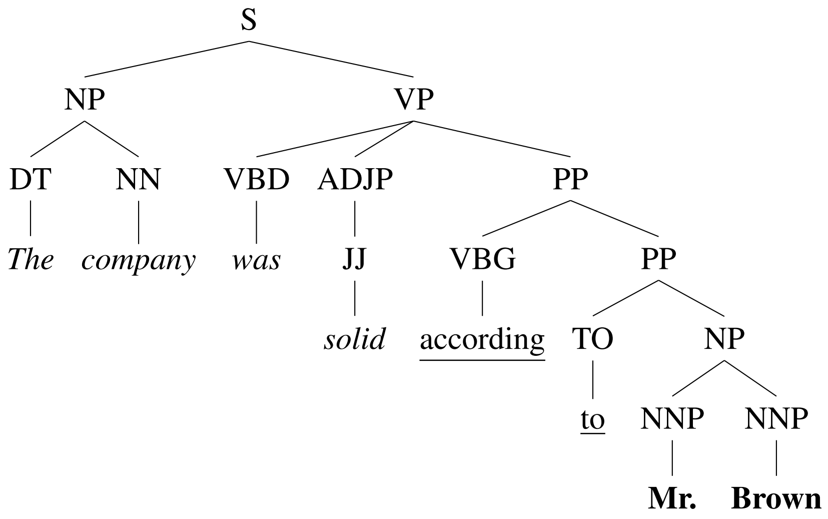


- Addition of sentence and verb features that account for nested verb-cues, identified as the verb-cues headed by another verb-cue. This was introduced to reduce the model errors due to nested attributions. These in some cases cause the system to label the first token of the nested content (e.g. ‘to’ in Ex. (99)) as the initial token of the content span, leading to an only partially correct prediction.

(99) **Charles Haworth, a lawer for Sunbelt**, SAYS *he* PLANS *to file a brief this week* URGING *the district judge to dismiss the suits, because Sunbelt’s liabilities exceeded its assets by about $2 billion when federal regulators closed it in August 1988.* (wsj_2354)

- Addition of sentence, verb and syntactic features for the prepositional cue ‘according to’, which is relatively productive and indicative of an AR but has a different syntactic structure with respect to cue-verbs. While for most cue-verbs the source is in subject position and the content is a dependent of the verb, for ‘according to’ the source is expressed as the object of the preposition that accompanies the verb ‘accord’ and the content as a prepositional modifier of the verb, as in the example below.





- Single-token content spans were excluded both from the gold and the predicted contents. These are commonly sentence-initial discourse connectives as in Ex. (100), followed by an attribution span and then a content span. These spans were annotated as part of the content since the annotator perceived the connective as part of the attribution.

(100) *Meanwhile*, analysts said *Pfizer’s recent string of lackluster quarterly performances continued, as earnings in the quarter were expected to decline by about 5%.* (wsj_2341)

- Leading and trailing commas and sentence punctuation were removed from each content span.

	Strict			Partial		
	P	*R*	*F*	*P*	*R*	*F*
B_{pun}	69	96	80	99*	97	98
Token	94	88	91	99*	93	96

Table 4.12: PARC 3.0 content span extraction results on direct ARs. The token based approach is trained and tested on all quotations. All differences between models are statistically significant, with $p<0.001$. (*) Precision is not directly comparable since the baseline is credited also for matching mixed ARs, while the model is scored relative to direct ARs only.

4.3.2.1 Results

This section presents the results on PARC 3.0 for the task of extracting the content spans of ARs, of which quotations represent a subset. Table 4.12 shows the results on direct ARs while Table 4.13 summarizes the results for indirect and mixed ARs and the overall results.

For direct ARs, the baseline achieves higher strict and partial *recall*, 96% and 97% respectively against 88% and 93% of the Token-based model. The model, however, has much higher strict *precision*: 94% while the baseline has only 69%. As discussed in Sec. 4.3.1.3 the baseline looses precision because it fails to recognize scare quotes and quoted titles, but also since it recognizes the quoted portion of mixed ARs as a direct AR.

Over indirect and mixed AR, the token-based model achieves much higher results than both lexical and syntactic baselines. Results are still relatively low, with the content span of indirect ARs being identified with 56% strict *recall* and 78% strict *precision* and mixed ones with 60% and 67% respectively. Over all ARs, the model is able to identify content spans with 71% strict *F-score*, and 82% partial *F-score*. The results show how the identification of an AR content span is a non trivial task. It cannot be successfully addressed by a strictly syntactic approach and remains an only partially solved problem. The following section presents an analysis of the model errors.

Concerning the syntactic baseline, this fails to recognise several types of ARs structures. In particular, since it identifies the content as the clausal complement of a verb-cue, the following ARs cannot be correctly retrieved:

- ARs having a cue other than a verb such as a noun or a preposition. Approximately 8% of ARs fall in this group (see Table 3.13). Since the model uses verb as well as noun cues and treats them as features, it is able to recognise also ARs having a noun cue or another type of cue.

- First-level ARs only. Nested ARs are also identified by the baseline if they are expressed by a verb-cue taking a clausal complement. While nested ARs are challenging also for the model, this can rely on features to recognise whether a verb-cue is part of a verb phrase headed by another verb-cue.

- ARs expressing the content span with a structure other than a clausal complement. AR content spans can also be commonly expressed by non-clausal elements such as NPs (approximately 8% of ARs) taking the role of a direct object or a passive subject. Some clausal content spans are expressed by structures that are not identified by the baseline, such as the case when the attribution span is expressed by a parenthetical within the content span (around 12% of ARs). The model can instead learn to associate the content span to different structures since its features comprise syntactic and dependency relations between each token in the sequence and a potential cue.

- Inter-sentential ARs. Since the syntactic baseline is constrained to the sentence boundaries, around 10% of ARs, whose content span extends over more than a sentence, cannot be correctly identified. The model does not have such constraint since the complete document is taken into account at labelling time.

4.3.2.2 Error Analysis

This section presents a systematic analysis of the main sources of error for the content extraction system. There are three possible type of errors: content boundaries, missed content and added content. Examples in this section identify the gold span with *italics* and the predicted span with **bold** font.
Content boundaries (i.e. gold and predicted spans overlap but are not the same):

- Nested ARs: the presence of a nested AR and in particular its cue (‘refused’ in Ex. (101)) can mislead the model. This identifies the nested content instead of the content of the first-level AR, leading to the identification of only part of the gold content.

	Indirect			Mixed			All		
Strict	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
B_{lex}	$42^{+\diamond}$	$31^{+\diamond}$	$36^{+\diamond}$	$48^{+\diamond}$	$39^{\diamond}$	$43^{+\diamond}$	$48^{+\diamond}$	$39^{+\diamond}$	$43^{+\diamond}$
B_{syn}	$70^{\bullet\diamond}$	$36^{\bullet\diamond}$	$48^{\bullet\diamond}$	$72^{\bullet\diamond}$	$45^{\diamond}$	$56^{\bullet}$	$74^{\bullet\diamond}$	$49^{\bullet\diamond}$	$59^{\bullet\diamond}$
Token	$78^{\bullet+}$	$56^{\bullet+}$	$65^{\bullet+}$	$67^{\bullet+}$	$60^{\bullet+}$	$63^{\bullet}$	$80^{\bullet+}$	$63^{\bullet+}$	$71^{\bullet+}$
Partial	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
B_{lex}	$65^{+\diamond}$	$52^{+\diamond}$	$58^{+\diamond}$	$81^{\diamond}$	$65^{\diamond}$	$72^{\diamond}$	$74^{+\diamond}$	$63^{+\diamond}$	$68^{+\diamond}$
B_{syn}	$90^{\bullet\diamond}$	$44^{\bullet\diamond}$	$59^{\bullet\diamond}$	96	$55^{\diamond}$	$70^{\diamond}$	$93^{\bullet\diamond}$	$58^{\bullet\diamond}$	$71^{\bullet\diamond}$
Token	$91^{\bullet+}$	$66^{\bullet+}$	$77^{\bullet+}$	$91^{\bullet}$	$81^{\bullet+}$	$86^{\bullet+}$	$93^{\bullet+}$	$73^{\bullet+}$	$82^{\bullet+}$

Table 4.13: PARC 3.0 content span extraction results. *All* reports the results over all quotations (direct, indirect and mixed). For the baselines, this is a combination of the strategy in B_{lex} or B_{syn} with the rules for direct quotations. ($^{\bullet}$: significantly different from B_{lex}; $^{+}$: significantly different from B_{syn}; $^{\diamond}$: significantly different from Token.)

(101) *Big investment banks refused* ***to step up to the plate to support the beleaguered floor traders by buying big blocks of stock***, traders say. (wsj_2300)

- Sentence boundaries: some ARs continue in a consecutive sentence without any explicit discourse connective nor quotation marks. These cases are ambiguous also for human annotators. Semantic understanding and world knowledge are usually needed to determine whether the sentence is still part of the content (because it is unlikely to be a personal addition of the writer) or not. The following are two examples of the above, Ex. (102) the model erroneously labelled the second sentence as a content span, in Ex. (103) the second sentence was labelled by the annotators as part of the content span, but not recognised by the model.

(102) a. ***On the exchange floor, "as soon as UAL stopped trading, we braced for a panic,"*** said one top floor trader. **Several traders could be seen shaking their heads when the news flashed. (wsj_2300)**

b. The maker of computer-data-storage products said ***net income rose to $4.8 million, or 23 cents a share, from year-earlier net of $1.1 million, or five cents a share***. *Revenue soared to $117 million from $81.5 million. (wsj_2332)*

- Attachment ambiguity: in particular coordinating conjunctions (Ex. (103)), but also some adverbial cue modifiers, can lead to multiple interpretations based on which one of the possible attachment readings is chosen. While human annotators can rely on higher levels of analysis, the model is bound to syntactic imprecisions.

(103) a. And Carl Spielvogel, chief executive officer of Saatchi's big Backer Spielvogel Bates advertising unit, said ***he had offered to lead a management buy-out of the company**, but was rebuffed by Charles Saatchi.* (wsj_2331)

b. A spokesman said ***the company's first quarter is historically soft**, and computer companies in general are experiencing slower sales.* (wsj_2342)

Missed contents (i.e. the gold span has no corresponding predicted span). This occurs more often with the following structures:

- Passives and impersonal structures.

(104) a. Also supporting prices are expectations *that the Soviet Union will place substantial buying orders over the next few months.* (wsj_2330)

b. There also are recurring reports *that the Soviet Union is having difficulties with its oil exports and that Nigeria has about reached its production limit and can't produce as much as it could sell.* (wsj_2330)

- Complex or less commonly occurring structures.

(105) a. The problem, however, is that GM's moves are coming at a time when UAW leaders are trying to silence dissidents who charge *the union is too passive in the face of GM layoffs.* (wsj_2338)

b. Time magazine executives predictably paint the circulation cut *as a show of strength and actually a benefit to advertisers.* (wsj_2350)

c. Against that backdrop, UAW Vice President Stephen P. Yokich, who recently became head of the union's GM department, issued a statement Friday blasting *GM's "flagrant insensitivity" toward union members.* (wsj_2338)

- Noun cues: these cues are less frequent and their relation with the content span is less syntactically encoded.

(106) a. Amid a crowd of crashing stocks, Relational Technology Inc.'s stock fell particularly hard Friday, dropping 23% because its problems were compounded by disclosure *of an unexpected loss for its fiscal first quarter*. (wsj_2342)

b. When the news broke *of an attempted coup in Panama* two weeks ago, Sen. Christopher Dodd called the State Department for a briefing. (wsj_2351)

Added contents (i.e. the predicted span does not correspond to a gold span):

- Void contents: structures resembling an AR but having no content expressed.

(107) Here's what several leading market experts and money managers say **about Friday's action, what happens next and what investors should do**. (wsj_2376)

- Semantic ambiguity: the same structure and cue of an AR may be used, however, with a non-attributional meaning. In Ex. (108), 'suggest' is used with its meaning of 'providing evidence or make someone think something' rather than that of 'making a suggestion'.

(108) That rise came on top of a 0.7% gain in August, and suggested **there is still healthy consumer demand in the economy**. (wsj_2358)

4.4 Source Extraction and Attribution

This section presents the system for the extraction of AR source spans. The system uses a model from the literature that attributes quotations to the entity speaker as its initial step. The model is adapted and used to attribute content spans to an entity mention. An algorithm is then applied to expand the mention to the complete source span.

Traditionally, attribution studies (see Sec. 2.2) have been concerned with the identification of the speaker of a quotation, intended as the actual entity that uttered the quotation. In this perspective, they had to address the issue of resolving pronouns and incomplete mentions and retrieving the representative mention in the chain, as well as disambiguating mentions to match them to actual real-world entities. In the following section, the arguments in favour of the identification of the source span, rather than the entity it refers to, will be presented.

For the identification of the source, the speaker attribution model developed in O'Keefe et al. (2012) was applied to extract the entity mention connected to the content. Instead of linking the entity mention in a coreference chain and disambiguating it to its world referent, I developed and applied another step to reconstruct the complete source span, including all relevant modifiers of the mention.

4.4.1 Source Entity vs. Source Span

Although ultimately a coreference resolution step is also necessary, the proposed approach to attribution considers entity identification alone not sufficient as it would miss relevant information. Sources are often identified in connection to the AR, in order to motivate why it is relevant to report their words, thoughts or intentions. Their local description can be extremely informative and even more relevant than disambiguating them and resolving them. Apart from well-known entities, many unknown entities (e.g. 'witnesses', 'one-time experts', ...) are mentioned in news. They are referred to with or without their proper name as this is not used to identify them, but simply to present the information as more accurate. These entities are identified by their description, usually in the form of an apposition or a relative clause as in Ex. (109). By turning attribution into a NE identification and resolution task, certain characteristics of the source that could affect the content, for example by showing the source's bias, expertise, attitude and relevance, would be missed.

(109) a. ***John Rowe****, president and chief executive officer of New England Electric,* said (wsj_0013)

b. ***Wilbur Ross Jr.*** *of Rothschild Inc., the financial adviser to the troubled company 's equity holders,* said (wsj_0013)

c. according to ***Mr. Cleveland****, a former UPS employee,* and others (wsj_1394)

d. says ***Sam Bridgers****, a neurologist who has studied the brain stimulators at Yale University* (wsj_0297)

e. predicts ***John J. Veatch Jr.****, an investment banker with Salomon Brothers who handled the Cowboys sale* (wsj_1411)

f. ***First Boston****, whose holding company, CS First Boston Group, is one of the larger issuers of bridge loans on Wall Street,* said (wsj_1415)

In several cases, such as in Ex. (110), coreference resolution would not give any additional information as the source cannot be linked to any NE. Sources of this kind are usually in the form of a plural common noun representing a profession which has relevant expertise or authority (e.g. researchers, analysts, government officials) or referring to a category of people (e.g. 'neighbours', 'witnesses', 'women') or even expressing a specific attitude or orientation.

(110) a. *doubters* say (wsj_2398)

b. *opponents* argued (wsj_0098)

c. *critics of poison pills* argue (wsj_0275)

d. even *some supporters* wonder (wsj_0765)

e. *democrats* argue (wsj_0343)

In other cases, what is relevant is not to identify the group, but to quantify the agreement within the group as in the examples in (111).

(111) a. ***some*** *analysts* and money managers think (wsj_1440)

b. ***several*** *executives* said (wsj_1447)

c. ***a few*** *experts*, going against the consensus, don't think (wsj_1623)

d. ***most*** *people* think (wsj_1617)

e. ***nearly half*** *of those who joined health clubs* said (wsj_0409)

f. ***60%*** *of the executives* said (wsj_0254)

Additional sources that do not have an accessible NE referent are also anonymous sources Ex. (112), studies and documents Ex. (113) and entities whose name is not known or considered not relevant Ex. (114).

(112) a. *one investment banker, who requested anonymity,* said (wsj_1822)

b. says *an official close to the case who asked not to be named* (wsj_0267)

(113) a. *a letter in the New England Journal of Medicine* notes (wsj_1825)

b. *a recent study for the Federal Aviation Administration* found (wsj_0730)

(114) a. says *a disembodied male voice* (wsj_0041)

b. recalled *one participant* (wsj_0745)

c. *one analyst* noted (wsj_0437)

d. *One takeover expert* noted (wsj_1305)

e. *a Coca-Cola spokesman* said (wsj_0245)

4.4.2 Models

The speaker attribution models described in joint work (O'Keefe et al., 2012; Pareti et al., 2013)[7] address the task of identifying the speaker of a quotation intended as the entity the quotation is attributed to. Although not coincident with the source span, one of the entity mentions corresponding to the quotation speaker is part of the source span. Since this is usually the head of the NP corresponding to the source span, its identification represents a good starting point for the retrieval of the complete span. Four models are defined:

1. Rule-based: the quotation speaker is selected by this method as the entity closest to the reporting verb nearest to the quotation. In case no reporting verb is found, it returns the entity closest to the end of the quote. Reporting verbs are identified using the list collected by Elson and McKeown (2010) and provided by the authors.

2. CRF: using an existing implementation from CRF-Suite (Okazaki, 2007) using maximum likelihood estimation with L2 regularization that chooses between up to 15 entities mentioned in the same paragraph of the quotation or in the ones preceding it.

3. NoSeq: a logistic regression implementation from LIBLINEAR (Fan et al., 2008) using maximum likelihood estimation with L2 regularization that outputs binary predictions (speaker vs. non speaker) for each candidate. Results are then reconciled by taking the candidate with highest probability.

4. Gold: the sequence model using gold labels for the previous predictions.

[7] The speaker attribution models were developed by Tim O'Keefe. I contributed to this task the data relative to PARC 2.0 and suggested some of the features.

Quotation speakers are annotated in the SMHC, together with the coreference chains including all the mentions of an entity. Coreference was not part of the PARC 2.0 annotation effort as the proposed approach considers it out of the scope of attribution. Entity candidates for this task were therefore taken from the BBN pronoun coreference and entity type corpus (Weischedel and Brunstein, 2005). This includes gold annotation of NEs as well as common noun and pronouns while only pronouns are resolved to their referent entity. Entities were taken as the speaker of a quotation if they were matching the beginning of a source span. The text for both corpora was encoded following Elson and McKeown (2010) by replacing all quotations, reporting verbs and speakers with a symbol. Reporting verbs were taken from the list provided by the authors which includes over 6,000 tokens. Sentences and paragraphs having no quotations and no mentions were removed as well as tokens tagged with a POS that was considered irrelevant, e.g. adjectives and adverbs.

Features

Features are calculated for each <quote(Q), speaker(S)> pair. They are a combination of positional, distance and frequency features, based on quotations, speaker mentions, speech verbs and punctuation appearing in the ten paragraphs preceding or including the quote. A set of sequence features is included and populated with either gold information or by using the predicted sequence of <Q, S> pairs. The set of features used is listed below, grouped by their type.

- Distance features: number of words/paragraphs/quotations/entity mentions between Q and S.
- Paragraph features: number of times S is mentioned and number of words and quotations in the paragraph including the quote and in the preceding 9 paragraphs.
- Nearby features: whether the tokens to the right or left of Q and S are punctuation/ another speaker mention/ another quote/ an identified speech verb.
- quotation features: whether S or other speakers are mentioned in Q; Q distance from the beginning of the paragraph; word length of Q.
- Sequence features: number of quotations attributed to S and number of quotations attributed to other speakers in the paragraph including Q and in up to 9 preceding paragraphs.

4.4.3 Source Span Identification

In order to retrieve the full span corresponding to a source (component (4b) in Fig. 4.1), that is the text span corresponding to the entity mention including local modifiers such as adjectives, quantifiers and appositives, I first applied the logistic regression model (NoSeq) to all attribution types and then extended the extracted mention to comprise all its modifiers. While I used the same baseline as Rule-based, using a list of reporting verbs was inadequate to identify other types of ARs. Instead of enlarging the list with other potential attributional verbs, I used the predictions from the verb-cue classifier (see Sec. 4.2.2). In order to overcome some shortcomings of the original implementation of the model, I also extended the implementation of NoSeq as described:

1. Only BBN entities of person type were added as mentions.

   The entity type list was extended to include additional BBN types. These correspond to:

   - ORG: organizations (companies, government agencies, institutions, ...).
   - GPE: geographical places (countries, cities, states, ...).
   - NOR: nationalities, religions, political.

2. Pronoun mentions were not linked to their anaphoric referent. Since only pronouns are resolved in the BBN, all entity chains consisted of only one mention. Retrieving the coreference chain of the entity is beyond the scope of this project since it is not needed in order to identify the AR source span and coreference resolution represents a separate task that can be addressed independently. For this reason, the single-mention 'chains' were considered suitable source head candidates for the attribution task and pronouns were kept as individual entities.

3. ARs sources were matched to BBN entities by taking as the gold entity the one having the same start as the source span. This matching resulted in a large proportion of source spans not having a corresponding entity. Entity-source span matching was firstly addressed by applying a simple algorithm which takes as the gold mention any entity mention overlapping with the source span. This raised the common issue of source spans containing multiple entity mentions. This issue could be partly addressed by naively taking as the gold mention the first one in the span. However, some sources in apposition constructions, such as in Ex. (115), would be incorrectly identified.

(115) the [organization] 's [founder] , [John Cannell] (wsj_0044)

Incorrect gold entity: organization

The adopted solution uses a hierarchy of entity types, taking as correct the overlapping entity higher on this scale:

$$person > organization > first\ entity$$

While this would be sufficient to correctly identify the gold entity in Ex. (115) it would still incorrectly identify the mention in the following source spans:

(116) a. former [Fannie Mae]'s [chairman], [David Maxwell]

Incorrect gold entity: Fannie Mae

b. Documents filed with [Gracenote]

Incorrect gold entity: Gracenote

Although imprecise, this is just a step towards the identification of the complete source span. The identification of an incorrect entity within the source span may still enable the retrieval of the correct source span.

4. Sources that did not match any entity, not having the same beginning, were added as entity mentions. All other overlapping mentions were discarded from the entity list. For source spans such as in Ex. (117) this would lead to the whole source span to be added as an entity, while all other entities part of the source span (shown in brackets) are removed.

(117) a. The [founder] of [Cailler], [François-Louis Cailler]

Candidate entities: the founder of cailler, François-Louis Cailler (Gold)

b. former [IBM] [founder] [Thomas J. Watson]

Candidate entities: IBM (Gold)

While the approach in point (3) considerably reduced the number of sources that could not be matched to a candidate entity, some cases of source spans matching no entity (e.g. ‘news’, ‘voice’) remained. To tackle this, the manually checked list of noun entities (Appendix C) that can appear as attribution cue as well as sources (e.g. ‘report’, ‘document’, ‘study’, ‘voice’) was used. All occurrences of these nouns in the corpus were added as candidate entities. Although not exhaustive, the list covers most of the cases, leaving only 1.8% of source spans in the Test set without a matching entity (Table 4.14). Even though all sources not matching an entity are counted as errors, it was decided not to add the source span itself to the entity list. A better solution would require developing an entity recognizer able to identify these entities that are common nouns.

5. ARs having no source are simply discarded by the model. In the Test set 8% of ARs have an implicit source (Table 4.14) and therefore no source span. Although some of these entities are syntactically represented by a Null subject trace and may be identified, this is beyond the scope of the current study. These ARs are discarded relative to the source span identification task, since there is no source span associated with them that the attribution component could retrieve.

6. Reporting verbs were identified through the list developed by Elson and McKeown (2010) and used to develop part of the features. Although extensive, the list focuses on reporting verbs and is therefore not adequate for the identification of different types of cue verbs. The predictions from the verb-cue classifier were therefore used.

Extending the entity span

Retrieving the source span was implemented as a subsequent step after the content was attributed to an entity. This was done using an algorithm (Algorithm 4) which:

- Identifies the head of the source span with the entity head, if this was a noun, or with the head of the entity parent otherwise.
- Takes as part of the source span all tokens in the sentence that are not part of the content and that have the identified head of the source span as an ancestor.

	TRAIN (0-22)	DEV (24)	TEST (23)	CORPUS
ARs no source	1019 (6.3%)	55 (8.8%)	93 (8.3%)	1167 (6.6%)
ARs no mention	369 (2.3%)	19 (3%)	20 (1.8%)	408 (2.3%)
Total ARs	15959	622	1111	17692

Table 4.14: Overview of the ARs in PARC 3.0 having no source span or to which no mention could be associated.

Algorithm 4 Source Span Identification

```
procedure IDENTIFYSOURCESPAN($AR$, predicted entity)
    if predicted entity is noun then
        $entityHead$ ← head of predicted entity
    else
        $entityHead$ ← head of the parent of predicted entity
    for all token in sentence do
        if token not in AR content and token has entityHead as ancestor then
            add token to AR source span
```

4.4.4 Results

Results for the entity attribution task and the source span identification are reported separately in the following sections.

4.4.4.1 Entity Attribution

Table 4.15[8] reports the results of the speaker attribution models presented in (Pareti et al., 2013). The data used for this task has been described in Sec. 4.1.1 and corresponds to PARC 2.0 and the SMHC corpus. Results are expressed in terms of accuracy since the model is forced to make a prediction for each AR content span and therefore the number of gold and predicted spans is equal.

Results for both corpora were considerably better using the logistic regression model (NoSeq), which reached an accuracy of 77% compared to 73% of the sequence model. I therefore adapted and applied the NoSeq model to the whole corpus and to all attribution types in PARC 3.0 and compared it to the simple Rule-based approach.

Accuracy results over the whole dataset are summarized in Table 4.16. The modifications to the baseline and the NoSeq model described in Sec. 4.4.3 had a positive impact on the results. The overall accuracy of the NoSeq model is 92% and of the Rule-based approach 75%. It is interesting to note the much lower score for indirect ARs (90%) with respect to direct ones (99%). A similar difference is found in the content span identification scores and can be ascribed to the lack of punctuation cues. Since punctuation does not play such role in the identification of the entity source, the scores suggests that indirect ARs have also more complex structures than direct ones.

[8] While I used an extension of the NoSeq implementation in further experiments, I did not recreate the entity attribution results reported in Pareti et al. (2013). Statistical significance tests are therefore not available for these results.

Corpus	Method	Dir.	Ind.	Mix.	All
PARC 2.0	Rule	70	60	47	62
	CRF	82	68	65	73
	NoSeq	85	74	65	77
	Gold	88	79	74	82
SMHC	Rule	89	76	78	84
	CRF	83	72	71	78
	NoSeq	91	79	81	87
	Gold	93	81	83	89

Table 4.15: Entity attribution accuracy results on all corpora over gold standard quotations/ARs.

Corpus	Method	Dir.	Ind.	Mix.	All
PARC 3.0	Rule†	90	72	64	75
	NoSeq†	98	90	97	92

Table 4.16: Entity attribution accuracy results on PARC 3.0 over gold standard quotations/ARs. †The Rule and NoSeq model with the modifications described in Sec. 4.4.3. Differences between Rule and NoSeq models for PARC 3.0 are statistically significant, with $p<0.0001$ (McNemar's test).

4.4.4.2 Source Span Identification

Entities identified by the NoSeq model were expanded to match the whole source span as described in Sec. 4.4.3. Strict and soft accuracy results are summarized in Table 4.17. This step uses the same baseline as for the entity attribution step and expands the entity using the same algorithm used for the NoSeq model.

Taking the entity identified by the NoSeq model and applying Algorithm 4 to include all the tokens in the mention as well as appositives and relative modifiers, yields satisfactory results. Source spans could be identified with 84% strict and 89% partial accuracy over gold standard AR contents. The increase with respect to the rule-based baseline is 11% strict and 12% partial accuracy.

Compared to the entity attribution step, soft results increased, while strict results decreased. This is explained by the fact that even for correctly identified entities, the identification of the full source span might be only partially correct. However, complex source spans may contain several entities. One of these entities might be identified, although not the correct one, leading to the source span to be partially correctly identified. The application of the algorithm to expand the retrieved entity to the complete source span may lead to the identification of a completely correct source span (Ex. (118)).

(118) That's when **George L. Ball, chairman of the Prudential Insurance Co. of America unit**, took to the internal intercom system to declare *that the plunge was only "mechanical."* (wsj_2300)

Gold entity: George L. Ball

Baseline entity: America

Gold source span: George L. Ball, chairman of the Prudential Insurance Co. of America unit

Baseline source span: George L. Ball, chairman of the Prudential Insurance Co. of America unit

4.4.4.3 Error Analysis

This sections presents an overview of the most common sources of error affecting the model identifying the source span. Examples identify the source gold span with bold font and the predicted one with small caps.

	Method	Dir.	Ind.	Mix.	All
strict	Rule	80	$72^{\bullet}$	$71^{\bullet}$	$73^{\bullet}$
	NoSeq	84	84^{+}	86^{+}	84^{+}
soft	Rule	97	$82^{\bullet}$	$87^{\bullet}$	$86^{\bullet}$
	NoSeq	99	93^{+}	98^{+}	95^{+}
partial	Rule	87	$74^{\bullet}$	$75^{\bullet}$	$77^{\bullet}$
	NoSeq	91	88^{+}	92^{+}	89^{+}

Table 4.17: Source span identification accuracy results for PARC 3.0 over gold standard AR contents. For the soft results, any overlap of the gold and predicted source spans was counted as a correct match. ($^{\bullet}$: significantly different from NoSeq; $^{+}$: significantly different from Rule).

- Incorrect entity or mention: an incorrect entity is identified leading to an incorrect source span. In some cases, even when the entity is part of the source span, the complete source span cannot be retrieved. When the wrong mention of a correct entity is identified an error occurs because no coreference data is available.

(119) a. *"You say you could have sold X percent of this product and Y percent of that,"* recalls **Theodore Semegran, an analyst at Shearson Lehman Hutton who went through this exercise during his former career as** A CHEMICAL ENGINEER. "And then you still have to negotiate." (wsj_2314)

b. **He**, like JUSTICE BRENNAN, considers *dissents highly important for the future*, a point that hasn't escaped legal scholars. (wsj_2347)

c. **The executive close to Saatchi&Saatchi** said *that "if a bidder came up with a ludicrously high offer, a crazy offer which Saatchi knew it could not beat, it would have no choice but to recommend it to shareholders. But (otherwise) it would undoubtedly come back "with an offer by management"*. THE EXECUTIVE said any buy-out would be led by the current board, whose chairman is Maurice Saatchi and whose strategic guiding force is believed to be Charles Saatchi. (wsj_2331)

- Parse errors: these are incorrect parse attachments leading to insertions or supplemental information to be included in the noun phrase.

(120) a. **A Poquet spokesman**, for example, criticizes *the Atari Portfolio because it requires three batteries while the Poquet needs only two.* (wsj_2387)

b. But **Mrs. Hills**, speaking at a breakfast meeting of the American Chamber of Commerce in Japan on Saturday, stressed *that the objective "is not to get definitive action by spring or summer, it is rather to have a blueprint for action".* (wsj_2321)

- Rules failed to recognize the NP and its modifiers.

(121) a. *"It seems to me that a story like this breaks just before every important Cocom meeting,"* said **a Washington lobbyist for a number of U.S.** computer companies. (wsj_2326)

b. *"It hasn't had any impact on us, nor do we expect it to,"* said a spokeswoman **for Miller Brewing Co., a major client of Backer Spielvogel**. (wsj_2331)

- Errors on the annotation side. In the example, *Yet* is incorrectly included in the gold source span.

(122) **Yet** more than one American official who sat in with her during three days of talks with Japanese officials said *her tone often was surprisingly "conciliatory."* (wsj_2321)

4.5 Extracting the Complete Attribution Relation

Sec. 4.3 and 4.4 described the models for the extraction of the content of an AR and its source span. Although the literature, in particular in the field of quotation extraction, would stop here, the AR is not yet complete. This section presents the final steps necessary to also identify the cue span of the AR (Sec. 4.5.1), completing the pipeline model and enabling the complete automatic extraction of all types of ARs.

4.5.1 Identifying and Linking the Cue Span

Cues are first identified by the model and used as features for the identification of content and source spans. However, content span extraction is the first step in my methodology for identifying an AR. Once the content span is identified, the second step attributes it to its source. In order to have the full AR, the cue span for that specific AR needs to be determined. This step was addressed by using a selection algorithm to identify the cue of an AR choosing among different candidate ones and expanding it to comprise the rest of the verb group and verb modifiers. The cue for a given content is selected with Algorithm 5 as follows:

1. The verb-cue head of the content span.

2. The verb-cue head of the source span.

3. The verb-cue closest to the source span, if available, or to the content span, that occurs in the same sentence.

4. The noun cue closest to the source span, if available, or to the content span, that occurs in the same sentence.

5. The token closest to the beginning or the end of the content span, within the same sentence.

Once identified, the cue is then expanded by adding all modal adverbials, auxiliaries, negation particles and phrasal verb particles that are descendants of the cue head.

Results for this step are summarized in Table 4.18. Results are calculated over all ARs as well as for each quote type individually. The algorithm that extracts the cue span is compared to a baseline that selects as the cue span the closest verb-cue to the content span. Assuming perfect choice of source and content spans, the AR cue span could be identified with strict accuracy of 90% and partial accuracy of 93%, both of which are over 10% more than the baseline. Also for this task, results are lower for indirect ARs, showing that not only their content span is harder to identify, but also the cues and AR structures connected to this quote status are more complex. A comparable analysis over non-gold source and content is presented in the context of an overall system analysis later in Sec. 4.5.3.

Algorithm 5 Cue Linking

```
procedure LINKCUE(AR)
    if AR content span has verb-cue as head then
        cue ← verb-cue that is head of content
    else if AR source span has verb-cue as head then
        cue ← verb-cue that is head of source
    else if AR sentence has verb-cue then
        if AR has source then
            cue ← verb-cue closest to source
        else
            cue ← verb-cue closest to content
    else if AR sentence has noun cue then
        if AR has source then
            cue ← noun cue closest to source
        else
            cue ← noun cue closest to content
    else
        if AR sentence has token before content then
            cue ← token before content
        else if AR sentence has token after content then
            cue ← token after content
```

	Method	Dir.	Ind.	Mix.	All
strict	Closest Cue	95	75•	83•	80•
	Algorithm	96	87+	92+	90+
soft	Closest Cue	96	78•	85•	82•
	Algorithm	96	92+	95+	93+

Table 4.18: Cue span identification accuracy results over gold standard AR source and content spans. For the soft results, any overlap of the gold and predicted cue spans was counted as a correct match. (•: significantly different from Algorithm; +: significantly different from Closest Cue)

4.5.1.1 Error Analysis

An incorrectly selected cue is often the result of an incorrect prediction at extraction time. The cue classifier might fail to identify a cue, thus leading to an incorrect match at linking time, or erroneously identify as a cue something that is not and that might be chosen by the linking algorithm (Algorithm 5) as the cue for that AR.

In Ex. (123), the incorrectly predicted (P) VERB-CUE gets selected before the gold (G) noun cue by the algorithm.

(123) G: In July, **the company** STUNNED Wall Street with the prediction *that growth in the personal computer business overall would be only 10% in 1990*, a modest increase when compared with the sizzling expansion of years past. (wsj_2365)

P: In July, **the company** STUNNED Wall Street with the prediction *that growth in the personal computer business overall would be only 10% in 1990, a modest increase when compared with the sizzling expansion of years past.* (wsj_2365)

Similarly, incorrect extraction of the content span may lead to incorrect identification of the cue. For example, if the cue has incorrectly been included in the content span, it cannot be found by cue identification. Alternatively, if nested cues have incorrectly been excluded from the content span, one of them may be selected by cue identification, again leading to an error. In Ex. (124), the incorrectly predicted (P) content span includes the gold(G) cue, thus discarding this as a cue candidate.

(124) G: Some analysts hedge their estimates for Quantum, because it isn't known *when the company will book certain one-time charges*. (wsj_2398)

P: **Some analysts** hedge their estimates for Quantum, *because it isn't known when the company will book certain one-time charges*. (wsj_2398)

4.5.2 Baseline

In order to evaluate the results on the full task, I developed an additional baseline that extracts the attribution triplet of source, cue, content. The baseline works similarly to the syntactic model proposed by de La Clergerie et al. (2009), which identifies the content of an AR as the grammatical object of a reporting verb and the source as its subject. The attributional verb is here added as the predicted cue. While their system uses a manually collected list of 114 verbs, the baseline can rely on the verb-cues recognized by the classifier.

For each identified verb-cue, the baseline identifies:

- content: the span corresponding to the clausal complement (*ccomp* relation) or the direct object (*dobj* relation) of the cue element
- source: the span corresponding to the subject (*subj* relation) of the cue element
- cue: the verb-cue element.

Each predicted AR is then matched to a gold one by looking for any gold AR having a content span overlapping with the predicted content span.

4.5.3 Results

While the previous sections have presented results on the individual components separately and independently from the other steps, this section will evaluate the complete AR extraction system on PARC 3.0, applying the source and cue span extraction models to predicted data. The steps are applied in the order described in Sec. 4.1.2. The order is important since it affects the error propagation. Other orderings of the steps would likely produce different results. Strict, soft and partial results for the individual tasks (over gold or predicted data) are summarized and compared to the baseline. The individual results are then aggregated with equal weighting providing the overall score for the complete AR extraction task.

Content span extraction results are summarized in Table 4.19. ARs extracted by the Token model described in Sec. 4.3.2 are compared to the final baseline. Content spans are identified with 71% strict and 82% partial *F-score*. Although this represents a large improvement over the baseline (51% and 66% respectively), the still relatively low results mean that errors in calculating source spans and cue spans from the content spans predicted in this step will propagate to the other tasks. The mean inter-annotator agreement score for the identification of the content span was 94% (see Sec. 3.4.2) which corresponds to an *F-score*[9] of 93%. The score was calculated over commonly identified ARs considering the span overlap similarly to the partial score and is therefore comparable to the partial *F-score*, which is 82% for this task.

[9] For the agreement, this was calculated by taking the annotations from one annotator as the gold ones and those from the other annotator as the predicted ones and using them to calculate the *F-score*.

	Strict			Soft			Partial		
CONTENT	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Baseline	59	45	51	82	63	72	79	56	66
Pipeline	80	64	71	95	76	85	92	74	82

Table 4.19: Content span extraction results for Strict, Soft and Partial metrics. The Pipeline model is compared against the Baseline. All differences between models are statistically significant, with $p<0.0001$ (McNemar's and Wilcoxon's tests).

The source span extraction component (Sec. 4.4.3) achieves relatively high result on gold content spans with 84% strict and 95% soft *F-score*. However, results drop to 71% and 79% respectively when the predicted content spans are used as shown in Table 4.20. This still represents a better score than the 62% strict and 68% soft *F-score* of the syntactic baseline. The mean inter-annotator agreement score for the identification of the source span was 91% (see Sec. 3.4.2) which corresponds to an *F-score* of 89%.

	Strict			Soft			Partial		
SOURCE	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Baseline	72	54	62	79	59	68	68	54	60
Pipeline	76	67	71	85	74	79	82	73	77
PipelineG	84	84	84	95	95	95	92	93	92

Table 4.20: Source span extraction results for Strict, Soft and Partial metrics. The Pipeline model is compared against the Baseline. PipelineG extracts a source span for each gold content spans in the corpus. All differences between models are statistically significant, with $p<0.0001$ (McNemar's and Wilcoxon's tests).

A similar detrimental effect due to error propagation can be observed for the cue span extraction component (Sec. 4.5.1). This is a downside of the proposed pipeline ordering, which tackles the hardest task first by starting with the content span identification. A different ordering leaving the content span identification last, would reduce the amount of propagation errors and would likely positively affect recall. It is foreseeable however to have a detrimental effect on precision for the reasons discussed in Sec. 4.1. Partial inter-annotator agreement for the identification of the cue span was close to 100% (see Sec. 3.4.2) both calculated as *agr* or *F-score*. Results, summarized in Table

4.21, show how the model for the identification of AR cues achieves also strict *F-score* results above 90%, however, using non gold content and source span data causes a drop of 19% *F-score* for strict and 14% for partial scores, while the baseline only reaches 57% strict and 61% partial *F-score*.

	Strict			Soft			Partial		
CUE	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Baseline	66	51	57	72	55	63	72	53	61
Pipeline	80	64	71	89	71	79	89	71	79
PipelineG	90	90	90	94	93	94	93	92	93

Table 4.21: Cue span extraction results for Strict, Soft and Partial metrics. The Pipeline model is compared against the Baseline. PipelineG extracts a cue span for each gold content span in the corpus. It uses the gold content and source spans to identify the span. All differences between models are statistically significant, with p<0.0001 (McNemar's and Wilcoxon's tests).

The individual components scores are aggregated with equal weighting and summarized in Table 4.22. Over the complete AR extraction task, the pipeline model identifies ARs with 71% strict and 83% partial *F-score*. When the components are run using gold data from the other components, the results increase to 82% strict and 88% partial *F-score*. The syntactic baseline, using the cues identified by the verb-cue classifier reaches a strict *F-score* of 57% and a partial one of 62%. The baseline results confirm once more the strong interconnection between attribution and syntax, while proving that syntax alone cannot fully represent this relation.

Results are also calculated for the extraction of ARs as a whole and reported in Table 4.23. Strict results report the accuracy for the identification of a completely correct AR, namely having completely correct source, cue and content spans. Soft results take as correct the predicted ARs whose content span overlaps with a gold content span. The pipeline model identifies complete ARs with 56% strict and 85% soft *F-score*, with an increase of 15% and 14% respectively over the baseline.

The results are promising, since they are close to the human agreement calculated for the further annotation of PARC 3.0 (see Sec. 3.4.2). We have to consider that the mean inter-annotator agreement for the identification of an AR, i.e. both annotators identified a completely or partially matching AR, was also 83%, calculated with the *agr* metric, and 80% *F-score*.

	Strict			Soft			Partial		
ARs	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Baseline	66	50	57	78	59	67	73	54	62
Pipeline	78	65	71	88	72	79	92	76	83
PipelineG	85	79	82	94	87	91	92	84	88

Table 4.22: Results on the extraction of ARs, i.e. source, cue and content spans. The Pipeline model is compared against the syntactic Baseline. Pipeline does not make use of gold data, predictions from one component are used as the input of subsequent components of the system. The results are calculated on source, cue and content individually and then recombined with equal weighting.

4.6 Conclusion

The system described in this chapter is a pipeline of different components that automatically extracts complete ARs, namely it identifies and links all its constitutive elements: source, cue and content.

The first step (Sec. 4.2) is a k-NN classifier that identifies AR cues by assigning binary labels to all head verbs. Verbs-cues are identified with 83% *precision* and 86% *recall*, showing a large improvement over list-based cue identification approaches, both in terms of *precision* and *recall*. Using the list of verbs adopted by Krestel et al. (2008) (Table 4.4) resulted in 74% *precision* and 70% *recall*. The cue classifier proved able to recognize unseen verbs and distinguish attributional from non-attributional uses of the same verb. For cues other than verbs (representing approx. 8% of the cases), the system relies on a list of potential noun cues and also identifies as cues all occurrences of ‘according to’.

Cues are then used as features in the second component of the system, a CRF sequence labeller that identifies which spans in a text correspond to the content of an AR. The model is first developed for assertion ARs only, which correspond to quotations. This enabled applying the approach to another corpus, the SMHC, and to compare the results to a baseline inspired by a common syntactic approach adopted in the literature. The syntactic approach identifies all quotations, including non-direct ones, as corresponding to the grammatical object of a verb-cue. Across all quote types (direct, indirect and mixed), results show a large increase in *recall* over the baseline, from 63% to 70% for strict evaluation and from 74% to 86% for partial, and a moderate loss in

	Method	*Precision*	*Recall*	*F-score*
strict	Baseline	47	36	41
	Pipeline	63	50	56
soft	Baseline	81	63	71
	Pipeline	97	76	85

Table 4.23: Source-cue-content complete triplet results. Strict results refer to ARs that are exactly identified. Soft score measure the partial identification of ARs, by taking as correct all ARs having a content span overlapping with the gold content span. Partial scores are not reported since they are not indicative, given the different token length of source, cue and content. All differences between models are statistically significant, with $p<0.0001$ (McNemar's and Wilcoxon's tests).

precision, from 80% to 76% for strict evaluation and from 92% to 87% for partial.

The CRF was then partly modified and retrained on all ARs from PARC 3.0 training set, comprising attribution types other than assertions. ARs in this dataset are more complex since it comprises a much larger proportion of indirect ARs. Indirect ARs lack punctuation clues and are therefore harder to identify. On this dataset, the model achieved 64% strict and 74% partial *recall* and 80% strict and 92% partial *precision*, well above the baseline results.

Content spans are the input of the source attribution component. The model used is a modification of the logistic regression speaker attribution model presented in O'Keefe et al. (2012) which makes a binary speaker/non-speaker classification of candidate entities for a given quotation and takes the highest probability one. The model identifies the correct entity with an overall accuracy of 91%. The identified entity is then expanded to comprise its modifiers by taking all tokens under the head element of its NP. Accuracy results for the identification of the source span are 84% for strict evaluation and 95% for soft.

Having identified content and source spans of an AR, the final step matches a cue to the content-source pair and expands it to comprise all modifiers that should be included in the cue span. Across all ARs, cue spans are identified with 90% strict and 93% soft accuracy.

Results are then recombined with equal weighting and compared to a syntactic baseline extracting the complete attribution triplet. The pipeline model is able to identify ARs reasonably well when using gold data to feed the different components, reaching 85% *precision* and 79% *recall* over strict matches. These results show the potential of the system, however they are an optimistic measure, since gold data would not be normally available. When run on predicted data, strict *precision* and *recall* drop to 78% and 65% respectively.

Overall, results are affected by the lower scores obtained on the content identification task, which remains the hardest and most crucial sub-task to tackle. Incorrectly identified content spans lead to incorrect sources being attributed and compromise the whole AR identification. Nonetheless, the pipeline model achieves better results that would be obtained using a syntactic approach, with an increase of 12% for strict *precision* and 15% for strict *recall*.

Finally, a measure of how well we can extract a complete AR extraction is given by scoring correctly or partially correctly identified source, cue and content triplets. Overall, 50% of ARs are exactly identified and 76% partially identified by the current methodology. Of the predictions the system makes, 63% are completely correct, while 97% do match at least in part an AR.

Chapter 5

Encoding Attribution in Other Languages and Genres

(Part of this chapter was published in Pareti and Prodanof (2010), Pareti (2015) and Cervone et al. (2014)) This chapter will present a contrastive analysis of the way attribution is encoded in languages other than English and genres other than news. This will provide a basis for discussing the portability and limitations of the current annotation scheme and look at attribution with a broader perspective.

In Sec. 5.1 I will draw on previous work on the annotation of attribution in a small corpus of Italian news texts. This will provide the basis to contrastively compare the way attribution is encoded in Italian and English.

In Sec. 5.2 I will present preliminary work, originating from the current approach to attribution, which extends the annotation to mailing list thread summaries and informal spoken dialogues. I will discuss the challenges of adapting the annotation scheme to these genres and present some of the differences between their encoding of attribution and the encoding of attribution in news.

5.1 English vs. Italian

The scheme for the annotation of ARs was initially applied to Italian news articles, leading to the creation of a corpus of 50 texts, the Italian Attribution Corpus (ItAC) (Pareti and Prodanof, 2010) (Sec. 2.1.3).

Attribution relations in Italian are expressed in a similar way as they are in English, thus the same scheme could be used for both languages. Unlike Italian, however, English can express attribution, to an unspecified source, by means of adverbials (e.g.

‘reportedly’, ‘allegedly’). These cases nonetheless fit the schema since sources can be left implicit as in Ex. (125).

(125) *Olivetti* reportedly *began shipping these tools in 1984.* (wsj_2326)

Unlike English, Italian can morphologically express evidentiality with mood, similarly to other European languages, such as French, Dutch and German. The use of the conditional mood as in Ex. (126b) does imply that the evidence supporting the statement is second hand. While this may imply that the statement is reported and it is therefore an attribution to an anonymous source, it can also just imply a lower confidence in the statement, namely an expression of epistemic modality.

It is not clear whether evidentiality entails attribution and the fact that the conditional mood can be used together with attribution as in Ex. (127) could imply that they are complementary but distinct. Moreover, for the devised approach to attribution, each AR should have a content span, an optional source span and a cue. The cue is not optional as it is needed to encode the type of attribution and the attitude the source holds towards the content. While the conditional mood in Italian can evoke attribution, it does not express a relation and is therefore not sufficient to establish an AR.

(126) a. L’incendio è:IND stato causato da una sigaretta.

b. L’incendio sarebbe:COND stato causato da una sigaretta.

The fire was caused by a cigarette.

(127) a. Secondo la polizia, l’incendio è:IND stato causato da una sigaretta.

b. Secondo la polizia, l’incendio sarebbe:COND stato causato da una sigaretta.

According to the police, the fire was caused by a cigarette.

Table 5.1 shows a comparison of the Italian pilot (ItAC) and the English PARC 3.0 AR corpora. Both corpora were annotated with the scheme developed for attribution. Although very different in size, some patterns already emerge. The comparison shows a smaller incidence of ARs per thousand tokens in the Italian corpus. This is more likely due to differences in style between the news corpora or to cultural differences rather than to characteristics of the language.

A much higher proportions of ARs in Italian (around 29%) do not have an associated source span. The proportion of ARs without a source in English is rather small (8%) and mostly due to passive constructions and other expressions concealing the source. These cases have usually been disregarded by attribution extraction studies focusing on the identification of the entity the source refers to, since they do not refer to a specific entity or they refer to an entity that is not possible to identify.

Italian, unlike English, is a pro-drop language, that is, subject pronouns are usually dropped because superfluous since a rich verb morphology already includes person-number information. The fact that in PARC 3.0 over 19% of source mentions are pronouns explains why Italian has around 20% more ARs without an explicit source than English. Unlike impersonal or missing AR sources in English, pro-drop sources in Italian usually refer to an entity and should be resolved.

	ItAC	PARC 3.0
Texts	50	2,280
Tokens	37k	1,139k
Toks/Text	740	500
ARs	461	19,712
ARs/text	9.2	8.6
ARs/1k tokens	12.5	17.3
ARs no source	29%	8%

Table 5.1: Comparison of AR news corpora of Italian (ItAC) and English (PARC 3.0) annotated with the AR scheme proposed in this thesis.

Some differences between the two languages also concern the choice and distribution of verb-cues. In a study comparing attribution in English and Italian opinion articles, Murphy (2005) noted that English commentators used more argumentative and debate seeking verbs while the Italian ones were more authoritative and consensus seeking.

By looking at the verb type distribution in the two corpora, it is worth noting the high proportion of attributional ‘say’ in English, around 50% of all cue verbs, which has no parallel in Italian. This might have to do with a tendency towards using a more neutral language in English as well as with repetitions and the use of broad meaning verbs being disapproved in Italian.

The annotation scheme for attribution could be successfully applied to both English and Italian, since they do not present major differences in the structures they use to express attribution. However, some agglutinative languages, such as Japanese, Korean and Turkish, can make use of verb suffixes and particles to mark what is reported and express reportative evidentiality. These languages would likely require some adaptation to the annotation scheme. For example, it should allow for a suffix to be annotated as cue.

5.2 News vs. Other Genres

While extremely frequent and relevant in news, attribution is not a prerogative of this genre. Very little work exists addressing attribution in other genres and it is almost exclusively limited to narrative (Sec. 2.1.2). In narrative, stylistic choices allow for a wider range of structures to be used, while sources are drawn from a small set of available characters. This section will present studies that take different genres into account and put the current annotation scheme to test.

PARC 3.0 already contains texts from different genres, albeit all related to news. The WSJ files included in the PDTB are classified into 5 different genres: essays, highlights, letters, errata and news. But what if we try to encode attribution in significantly different genres and we take into account other registers and domains? In order to test this, I will present here two preliminary studies inspired by the work in this thesis. In these studies attribution was annotated on two significantly different corpora: technical mailing thread summaries and informal spoken dialogues. A comparison of the corpora characteristics is reported in Table 5.2.

	PARC 3.0	SARC	KT-pilot
Genre	News	Dialogue	Thread summaries
Register	Formal	Informal	Informal
Medium	Written	Oral	Written
Tokens	1,139k	16k,2h	75k
ARs	19,712	223	1,766
ARs/1k tokens	9.2	14	23

Table 5.2: Comparison of AR corpora from different genres annotated with the AR scheme proposed in this work.

5.2.1 Attribution in Mailing Thread Summaries

The annotation schema for ARs was applied by Bracchi (2014) to a pilot corpus of mailing thread summaries (KT-pilot) sampled from the Kernel Traffic Summaries of the Linux Kernel Mailing List.[1] The corpus differs not only in genre, but also in register and domain. The summaries report what different people contributed to the discussion. A discussion consists in a back and forth of comments and replies. The register is rather informal and the domain is technical and related to computer science.

This corpus is particularly interesting for attribution since its language is significantly different from that used in news but also extremely rich in ARs. The corpus was studied by Duboue (2012), who investigated the various ways of reporting that could be used in summaries. Quotation-introducing verbs were automatically extracted by taking the past tense verb closest to the quotation span. Quotation contents are already marked in the corpus. The study identified 39 classes of verbs that introduce a quotation. It also recognised as not neutral a third of the types, mostly low-frequency, since they express an evaluation or convey the author or source's emotions.

While the annotation schema was suitable to encode ARs in this genre and did not require modifications, some differences emerged with respect to news texts. Bracchi (2014) reports preliminary analysis concerning the attribution cues. She identifies some characteristics of AR cues in the KT-pilot, for example the use of acronyms as attribution spans, e.g. IMHO: 'in my humble opinion' (Ex. (128)), AFAIK: 'as far as I know', IMNSHO: 'in my not so humble opinion'. These are not only unlikely to be found in news, but also combine together both source and cue. Since the annotation scheme allows for the source and cue element to overlap, these cases can be annotated by marking the acronym as corresponding both to the **source** and to the cue span (Ex. (128)).

(128) *This* **IMHO** *is a good thing for all Real Time SMP.* (Bracchi, 2014)

As Bracchi (2014) notes, the occurrence of attributional verb-cues in the KT-pilot is also more distributed, with 'say' covering only 18% of the cases (compared to around 50% in PARC 3.0) and almost 11% being covered by 'reply', a common verb in the mailing thread summaries but rather low-frequency in news. Moreover, the study identifies some common verbs strongly associated with attribution in news language, such as 'declare' and 'support', as exhibiting a preferred other use in the computer domain (e.g. 'declare a variable', 'support a version').

[1] Accessible at: http://kt.earth.li/kernel-traffic/archives.html

5.2.2 Attribution in Spoken Dialogues

Cervone et al. (2014),[2] investigate attribution in spoken informal telephone dialogues and explore the possibility to apply the proposed annotation scheme to a genre using a different medium of communication. The preliminary corpus (Speech Attribution Relation Corpus (SARC)) was annotated by a single annotator with a modification of the scheme for attribution. The same scheme, with source, cue and content elements, could also be applied to dialogues, although with the addition of the 'fading out' category. This category is borrowed from Bolden (2004) to account for additional words whose inclusion in the content is ambiguous. In Ex. (129), the part of the content span delimited by square brackets is considered as fading out, since it is uncertain whether it still is part of what was originally uttered.

(129) **I** told him *that I cared a lot about him [because I mean I've always been there for him haven't I]*

Although typical of the spoken medium, where only the beginning of a source shift is signalled by the speaker, 'fading out' has a parallel in written texts, where syntactic ambiguities can leave the content boundaries unclear as in the bracketed portion of the content in Ex. (130) which could be part of what the workers described as well as a remark the author adds. In PARC 3.0, it was up to the annotators to determine the boundaries of the content for each case, although indication was given as to adopt a minimal approach, thus excluding the ambiguous parts.

(130) **Workers** described *"clouds of blue dust" that hung over parts of the factory*, [even though exhaust fans ventilated the area]. (wsj_0003)

In SARC the relation between the speaker and each turn in the dialogue is not annotated as an AR. While dialogue turns in fiction or in news interviews would be ARs, turns in spoken dialogues are not. Their exclusion is motivated by the relation being not linguistically expressed. It is obvious to the participants in the dialogue who is the speaker of each turn, in this case the voice on one side of the line. The attribution of the turn to its speaker is not annotated since it is meta-textual or extra-textual. This treatment is consistent with the approach adopted in news, where the article attribution to its writer is not annotated.

[2] I contributed to this work with the preliminary idea of analysing attribution in speech and the initial annotation scheme; I also jointly worked on the contrastive analysis of the textual aspects of attribution in speech vs. news.

Some smaller differences with respect to news derive from SARC being a corpus of spoken and colloquial language. Apart from the use of colloquial attributional expressions such as ‘I’m like’ or ‘she goes’, that are not likely to appear in news, there are frequent repetitions and broken sentences. In Ex. (131), the source and cue of the AR are repeated twice. In news language, this would normally correspond to a case of nested ARs (i.e. Ellie just said to me yesterday: “She said: ‘Oh I’m a bit bored of the snow now mum’”). However, in the example there is only one AR and since an ARs should have only one cue, only the cue closest to the content is annotated. While an AR can have multiple sources, this is intended to represent the case when a content is attributed to more than one source (e.g. ‘toy manufacturers and other industrialists’) and not twice to the same source. The first source-cue pair is not an AR, since it is not complete as it lacks the content. Before this was uttered, the speaker interrupted the sentence and produced a new one, partly overlapping with the previous, but distinct.

(131) haven’t ye ah God do you know I was just off it now and **Ellie** just <u>said</u> to me yesterday **she** <u>said</u> *oh I’m a bit bored of the snow now mum*

The application of a lexicalized approach to attribution to the spoken medium proved more problematic. In particular, speech lacks punctuation, which instead plays a crucial role in written texts, allowing the identification of direct quotations and in some cases being the only lexical cue of an AR. In speech dialogues instead, part of the role played by punctuation is taken over by acoustic features. The preliminary analysis reported by Cervone (2014) shows some correlation of acoustic aspects, such as pauses, intensity and pitch, with the content boundaries. In the examples below (Cervone, 2014, p.102), acoustic features allow to reconstruct the ARs in the dialogue turn in Ex. (132a) as it is shown, with the help of punctuation, in Ex. (132b).

Moreover, not only the content boundary is defined by extra-textual clues, but in certain cases, the whole AR is reduced in the text to its content element. In spoken language, cues might be expressed by acoustic features and thus not identifiable from the text alone. In Ex. (132), “what for a loft” and “I’m not going to do that” are attributed to a different source (mentioned at the beginning of the turn as ‘she’). However, the source is left implicit and the cue replaced by acoustic means. The source shift is also suggested by the ‘and I’ sequence introducing the speaker’s own words.

(132) a. she wouldn’t I said well but I said at the end of the day I said you could sell your house what for a loft and I said well yes if you really didn’t have

any money you'd have to sell it for a loft buy something smaller well I'm not going to do that and I thought well then you haven't not got any money then have you it's not really the same thing

b. She wouldn't. I said: "Well but", I said: "At the end of the day", I said: "You could sell your house." "What? For a loft?" And I said: "Well, yes! If you really didn't have any money you'd have to sell it for a loft. Buy something smaller." "Well I'm not going to do that." And I thought: "Well, then you haven't not got any money then, have you?" It's not really the same thing.

5.2.3 Other Forms of Attribution

Not only in the spoken medium, but also in the web one, attribution can be expressed in extra-textual ways, thus requiring a partly different encoding. For example, attribution can rely on hypertext, both to express the source and to delimit the content span, e.g. by embedding in it a link to its source.

In addition, the web can make use of graphical elements to show the source of some text, e.g. by embedding part of another page or showing a tweet as an image. Attribution is also graphically expressed in the comics medium, where sources are drawn and cues are rendered by bubbles enclosing the text as in Fig. 5.1. The type of attitude is encoded by means of specific shapes of the bubble and by varying the line thickness or continuity.

Also in academic writing attribution is expressed in a distinct way, with sources being papers, commonly referenced in a strictly encoded way.

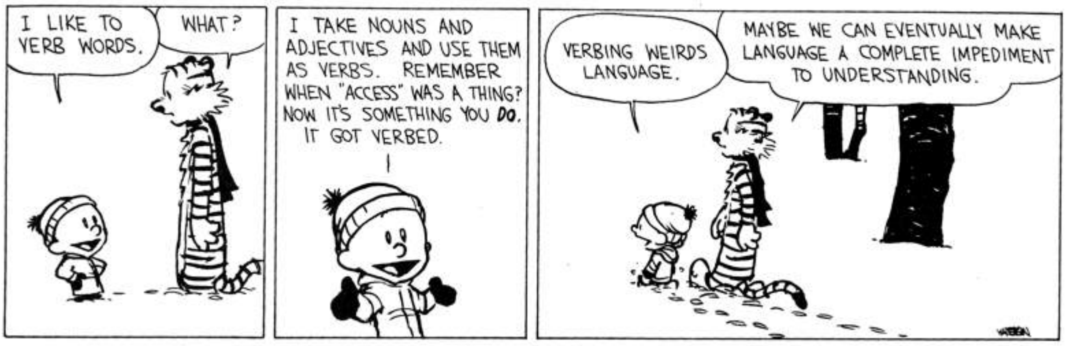

I LIKE TO VERB WORDS.
WHAT?
I TAKE NOUNS AND ADJECTIVES AND USE THEM AS VERBS. REMEMBER WHEN "ACCESS" WAS A THING? NOW IT'S SOMETHING YOU DO. IT GOT VERBED.
VERBING WEIRDS LANGUAGE.
MAYBE WE CAN EVENTUALLY MAKE LANGUAGE A COMPLETE IMPEDIMENT TO UNDERSTANDING.
WATSON

Figure 5.1: Example of attribution in comics (Watterson, 1994).

5.2.4 Conclusion

Different languages can resort to a range of different structures to encode ARs, including morphological markers of reportative evidentiality, which could represent a challenge to the current annotation scheme. The scheme could be applied to both Italian and English news texts without modifications. While some stylistic differences emerged, similar structures are used by both languages to express attribution, albeit with different distribution. The only structural difference that was identified is the use of attributive adverbials in English but not in Italian.

Attribution can be found in many types of human communication, whether verbal or not. Different attributive structures and means are used in different genres, including graphical and acoustic clues. I reviewed the application of the proposed lexicalised approach to attribution to informal and technical mailing list thread summaries and to informal telephone spoken dialogues. While the former presented only some distributional differences and the presence of additional attributional structures (i.e. acronyms), the latter required also some adaptation of the scheme. The main challenge to the annotation of ARs in speech is the need to also account for non lexical cues. Acoustic clues contribute and in some cases replace textual attributional elements.

Overall, preliminary applications of the current annotation scheme beyond English news texts showed good flexibility and coverage of the current approach. Nonetheless, some adaptation to different language structures and to different genres may be needed.

Chapter 6

Future Work and Potential Applications

This chapter describes some ideas for improving the identification of ARs as well as some potential applications.

In Sec. 6.1, I will discuss the possibility of adopting a joint approach to the extraction of ARs which could better exploit the interdependencies and strong connections among source, cue and content. In Sec. 6.2 I will then present some possible approaches to the identification of nested ARs.

Sec. 6.3 will then present preliminary investigations towards exploiting aspects of attribution to select information. I will explore the possibility of clustering sources into types, such as anonymous, experts and well-known, or deriving different degrees of reliability of the information conveyed by the content. Reliability is affected by factuality and evidentiality clues as well as the level of nesting (or embedding) of an AR.

An idea for a news summarization model based on attribution will be included and presented in Sec. 6.4. This will show how attribution could enable summarizing different viewpoints, i.e. attributions to different sources, as well as statements from the same source.

6.1 A Joint Model for Labelling Attribution Relations

An extension of the current work would be the joint modelling of attribution extraction. This would allow comparing a joint model to the proposed pipeline one. The pipeline model has the advantage of keeping the different components distinct. Components are individually optimized and their potential can be fully measured by evaluating as single component at a time while retrieving the other components from gold data. Another advantage of the pipeline approach is that models are independent and can be independently applied, or replaced, according to which task we want to address. For example, for direct quotations we might only be interested in applying the source attribution component.

On the other hand, a joint model would be easier to maintain and not be affected by propagation errors, which represent instead a limitation of pipeline models. Moreover, the different elements of an AR are interconnected and their identification is strongly dependent on the other elements, in particular for the source and content spans. This interconnection was captured in the pipeline model by making available at each step information related to the other AR elements, which could be used to derive additional features.

In order to enable this, we either made use of the already predicted elements or had to resort to a previously identified set of potential elements, such as potential cues and source entities. For example, when extracting the source, we already had its previously extracted content and could make use of the potential cues identified by the verb cue classifier. A joint model could better represent these interdependencies among source, cue and content by jointly extracting and linking them.

6.2 Detecting Nested Attribution Relations

Although their extraction has yet to be addressed in the literature, nested ARs are not a rare phenomenon. From the statistics on PARC 3.0 (see Sec. 3.5.4.1), it can be estimated that over 20% of ARs might be nested. Such proportion shows that this aspect of attribution has been underestimated and should be taken into account. While not explicitly addressed in this thesis, this section will present some approaches to the identification of nested ARs.

The same pipeline model developed in this thesis and presented in Ch. 4 could be recursively applied to the identified content spans of first-level ARs to identify nested ARs. However, better results can be expected from training a similar model on nested ARs only, considering each content span of first-level ARs as the document window. A dedicated model, with nested-specific features, would be able to address some of the peculiarities and challenges of nested ARs (see Sec. 3.5), in particular:

- Direct and mixed ARs and punctuation clues, in particularly quotation marks, are almost absent.

- Cues and attributional structures are slightly different or have a different distribution.

- There is a much higher incidence of non-assertion ARs.

- The length of source and content spans is shorter with respect to first-level ARs (the average length in tokens is 1.6 vs. 3.7 and 11 vs. 19 respectively).

- Sources are pronominal in 46% of nested ARs vs. 19% of first-level ones. Moreover, a third of these pronouns are in the first person, a phenomenon which is almost absent from first-level ARs. In addition, implicit sources are almost twice as frequent as in first-level ARs (13% vs. 7.7%).

- A higher proportion of cues is not a verb (14% vs. 8%).

Another possibility would be to rearrange the proposed pipeline in order to enable addressing the extraction of first-level as well as nested ARs at once. The proposed ordering starts with the identification of the content span, performed as a sequence labelling steps which does not allow for the identification of a sequence within a sequence. The extraction could instead start from the identification of the cue span. For each identified cue, which would include nested AR cues, we could then apply a content span identification step.

6.3 Attribution for Information Diversification[1]

The automatic extraction of ARs enables the development of applications such as source-sensitive information extraction, i.e. the possibility of extracting information together with its source. ARs could allow diversifying information by retrieving information about a topic expressed by different sources or different types of sources. Attribution could also allow excluding sources that are not considered trustworthy.

According to a quotation attributed to Ronald Harold Nessen, a former White House Press Secretary: *"Nobody believes the official spokesman... but everybody trusts an unidentified source."* Whether a provocation or an observation, it is undeniable that the source has a deep impact on the information conveyed by the attributed content. It

[1] Part of this section was published in Pareti (2011).

affects not only whether we believe or not what the content expresses, but also the way we interpret it. For example, Bernardini and Prodanof (2014) applied attribution to irony, suggesting that the very same message can be interpreted as ironic or not based on its source, thus opening a new direction for irony detection studies, traditionally focusing on the identification of clues of irony within the message itself.

Since by extracting ARs we have identified attributed text together with their source and cue, we can now identify those elements that might affect the information conveyed by the content. The possibility to distinguish categories of sources would enable the selection (or exclusion) not just of specific sources but also of source types (e.g. anonymous sources, hearsays) during extraction. Although considerable world knowledge may be needed to determine a source's degree of expertise, reliability and bias with respect to a specific matter, a coarse-grained classification of attribution sources from contextual information could be linguistically derived.

In particular, the source mention (i.e. the entity mention including modifiers, appositives and relative clauses) can be analysed to derive features that can be exploited to determine classes of sources. Examples of such features are: the use of a definite or indefinite article; the number (singular or plural); the presence of a description associated with the source first mention and whether the source itself is a NE and the presence of quantifiers.

An example of some possible source distinctions, inspired by the classification of discourse social actors presented by van Leeuwen (2008), is reported in Table 6.1. Types are shown together with an example and an indication of the features allowing their identification. Further investigations of the source mention and associated description would allow to refine the classification. In addition to a coarse classification based on features in context, the extracted sources could be matched to an external database such as Freebase[2] to derive more fine-grained classifications (e.g. profession and gender).

The source is not the only element of the AR affecting the content. The cue plays also a major role by suggesting the factuality or evidentiality of the AR and its content or by conveying the authorial stance. A classification of verb and noun cues may partially rely on ontological categories using existing resources such as VerbNet or WordNet. However, a more accurate analysis requires to take the complete cue span into account. In particular, adverbs expressing polarity, such as 'never' and 'not', attitude and mood, such as 'derisively', 'hesitantly' and 'happily'.

[2] www.freebase.com

SOURCE TYPE	EXAMPLE	SOURCE FEATURES
group	*scientists*	plural, mass noun
individual	*a man*	singular
named	*Mr. Wilson*	NE
unnamed	*a spokesman*	not NE
well-known	*Obama*	no description
not well-known	*Hajime Sasaki, an NEC vice president*	description
specific	*Mr. Wilson, corporate secretary*	def. or no article
one of many	*Nobuyuki Arai, an economist*	indef. article
collectivity	*analysts*	plural
aggregation	*some entrepreneurs*	plural with quantifiers

Table 6.1: Features for the identification of source types.

Finally, the level of nesting of an AR has an important impact on the content since it makes overt the attribution chain it is embedded into. This gives a measure of the different passages a content underwent, which might have caused manipulations of the original content. Thus, the more nested a content is, the less reliable it is. Moreover, the reliability of each embedding AR source will propagate to the embedded ARs. In Ex. (133), the nested content depends not only on the source 'Mr. Masson', but also on the source of the embedding AR, 'Ms. Malcom'. Thus, we should consider reliability and bias of both when judging the embedded content. The level of nesting may be used to determine a threshold below which we do not want to retrieve information, e.g. we might decide it should only be first-hand.

(133) **Ms. Malcolm**, for example, wrote [*that **Mr. Masson** described himself as ["the greatest analyst who ever lived."]]* (wsj_0944)

6.4 Attribution for Summarisation

News reports regularly feature people's views on topics of interest such as trials, earthquakes and political decisions. Automatically producing structured summaries of this information would allow views to be tracked both longitudinally (over time) and latitudinally (over view holders). However, standard automated multi-document summarisation is designed to collect and integrate facts based on significance (computed from lexical specificity) and novelty (computed from lexical dissimilarity). This is unsuitable for summarising quotations and opinions: lexical similarity might express the viewpoint of different agents on a same topic/facet as in Ex. (134). And even apparent synonyms may reflect vastly different perspectives as Klebanov et al. (2010) show. If we consider the use of 'feticide' as a synonym for 'abortion' in Ex. (134), we can see that it entails an entirely different viewpoint.

(134) a. The president declared: the government will not change the abortion law.

b. Sen. Brown fears that the government won't allow changes to the feticide law.

A three-step approach to solving the task of summarising viewpoints would be:

- Step 1: clustering quotation and opinion ARs, namely all quotations alternatives originating from the same quotation.

  Once ARs are extracted the similarity of attribution contents can be calculated using quote-clustering techniques, e.g. edit distance and word similarity (Leskovec et al., 2009), allowing however more variation for indirect quotations and opinions. ARs having not only high similarity but also the same source and a similar cue can be clustered together. Finally, all clusters attributed to the same source (or source group e.g. Liberals) can be grouped.

- Step 2: selecting relevant instances within and outwith clusters.

  From each cluster one AR instance should be selected, based on one of two strategies (depending on task):

  - Completeness (i.e. having the most complete span).
  - Salience (i.e. having the most repeated span).

  Within-cluster relevance can be scored by applying similarity detection not to sentences, but to the attribution contents of different clusters, assigning higher relevance to clusters having low similarity and a larger number of instances (i.e. that were reported by multiple articles).

- Step 3: generating a structured summary.

  Sentences containing relevant ARs can be grouped by source (i.e. all unique ARs from a source), by sub-topic or by perspective (e.g. pro-Life vs. pro-Choice).

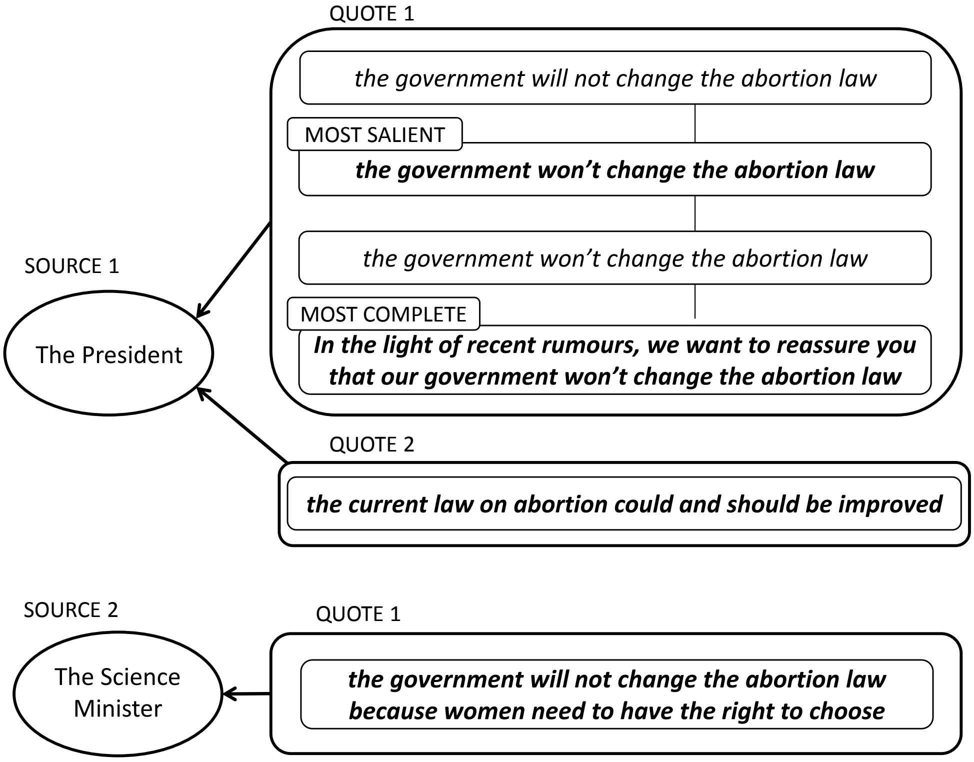

QUOTE 1
the government will not change the abortion law
MOST SALIENT
the government won't change the abortion law
SOURCE 1
the government won't change the abortion law
The President
MOST COMPLETE
In the light of recent rumours, we want to reassure you that our government won't change the abortion law
QUOTE 2
the current law on abortion could and should be improved
SOURCE 2
QUOTE 1
The Science Minister
the government will not change the abortion law because women need to have the right to choose

Figure 6.1: Example of attribution clustering (Step 1) and relevance detection (Step 2).

Attribution clustering would help overcoming the limitations of current quote-clustering systems such as Meme-tracking (Leskovec et al., 2009). Because these clustering systems are based only on word sequence and edit distance, the example from Source 2 in Fig. 6.1 may be erroneously considered as being drawn from the same original quotation as the examples in the Quote 1 cluster of Source 1. The proposed system, grouping ARs not only sharing similar words but also the same source may considerably increase quotation detection and clustering precision.

The inclusion of AR extraction allows identifying and summarising viewpoints in a cluster of articles and retaining similar sentences belonging to different sources. It would also enable novel strategies, such as summarising sources' statements and different perspectives.

Chapter 7

Conclusion

This thesis has proposed a computational approach to attribution that addresses the challenge of automatically extracting attribution relations in news texts.

Attribution is ubiquitous in news, and being able to reliably extract it is particularly important as it enables retrieving attributions to a specific entity, profiling entities and differentiating information by provenance. It also enables evaluating the quality of information by taking its source into account as well as by identifying whether information if first or second-hand.

Although attribution is receiving increasing attention in the literature, different fields are independently looking at a limited portion of the relation, in particular since it has relevant implications for quoted speech, discourse, opinion and factuality studies. Moreover, studies are mostly small-scale, lacking a large annotated resource that could serve as a common testing ground and make results meaningful and comparable. With only small and partially annotated corpora available, studies extracting subsets of attribution typically have to resort to heuristic models. Some of these approaches suffer from false assumptions that are not statistically grounded, for example that attribution is always a syntactic relation and sources are Named Entities.

This thesis focused on ARs independently from other linguistic levels and relations with which it interacts. The goal was to reach a deeper understanding of this relation and the ways it is encoded, in order to improve its automatic extraction. As a first step, I have adopted a more comprehensive approach to attribution, inspired by the range of ARs included in the PDTB. The approach (Ch. 1) includes quotation and opinion ARs, traditionally separately studied, as well as other types of attributive relations, and takes a wide number of attribution-bearing structures into account. One of the key ideas is to recast attribution extraction as the task of identifying the text spans encoding its source, cue and content elements.

Within this framework, I have created PARC 3.0 (Ch. 3), a large and fully annotated corpus of ARs. The annotation scheme was tested with an inter-annotator agreement study showing satisfactory results for the identification of ARs and high agreement on the selection of the text spans corresponding to source, cue and content. I have used the corpus, which comprises around 20k ARs, to investigate the range of structures that can be used to carry attribution. The results show a complex and varied relation of which the literature has addressed only a portion. PARC 3.0 can be used in a range of different studies to analyse attribution and validate assumptions as well as to develop supervised attribution extraction models.

This thesis contributes a complete system for the automatic extraction of ARs, described in Ch. 4. This is a pipeline of supervised models developed from the annotation in PARC 3.0. The system can identify and link source, cue and content spans of an AR with significantly higher precision and recall than traditional syntactic and rule-based approaches. This allows us to take fresh news texts and automatically identify different types of ARs in it, whether opinions, quotations or other types. We can not only connect the attributed text to its source, but also know the textual anchor of the relation. This is a relevant element that characterizes the relation by determining its type, factuality and evidential value and by carrying the source attitude and the authorial stance.

Apart from enabling the development of the attribution extraction system proposed in this thesis, PARC 3.0 has already allowed reaching a deeper understanding of the encoding of ARs in news. From the statistical analysis on the corpus and the results of the experiments on the extraction, we now know that:

- A significant proportion of ARs have no explicit source. In such cases the attribution might still link the content to a specific source that can be retrieved from the text but has no corresponding span in the relation, but also simply signify that the author takes the distance from what is expressed in the content by mentioning that it originates from a third party. While the quotation attribution literature starts from the assumption that all quotations have a source and address the task as a speaker attribution task, this approach is not suitable for a relatively small number of ARs.

- The majority of ARs are not delimited by quotation marks, thus their identification cannot be taken for granted. Identifying content spans and their boundaries for indirect and mixed ARs actually constitutes the hardest challenge for AR extraction.

- ARs are a more complex phenomenon than it appeared from the literature. They are not simply a syntactic phenomenon. This is clear just by considering that around 8% of ARs are inter-sentential. Moreover, ARs are expressed by a large range of structures. While a relatively large number of ARs is encoded by a few syntactic structures that are highly predictive of attribution, the remaining is expressed by a variety of structures that cannot be strictly encoded. Therefore syntactic approaches to the extraction of ARs lead to systems that are relatively precise on a subset of ARs, but have rather low recall.

- Although disregarded by the literature, nested ARs are a large proportion of attributions in news, where even more than 20% of ARs may be nested. Nesting is not just a recursive aspect of attribution, this subset of the relation presents its own peculiarities and less typical encoding with respect to first-level ARs, making it the hardest type of ARs to identify. Nested ARs are very rarely direct, are mostly not assertions and have a larger proportion of pronominal and implicit sources.

- Attribution has been studied in different linguistic areas, however, there is no exact overlap of attribution for any of them. Attribution cannot be easily reduced to a syntactic or discourse phenomenon. It does show strong interconnections with other levels of linguistic analysis and it has important implications for factuality and opinion studies, however, it remains a separate task.

- Some of the assumptions at the basis of several approaches in the literature are not confirmed by the data; in particular, the assumptions that content spans are clausal elements, sources are NEs and cues are verbs. While these are frequent cases, the corpus shows that a relevant proportion of ARs does not fit these constraints.

While the current encoding of attribution is rather comprehensive, some additions would be desirable. In particular, it would be useful for the annotation to also encode the entity the source refers to. This would enable supporting entity resolution for the source, which is a crucial step for opinion and quotation attribution studies. For opinion studies it would be relevant to also annotate the target of the opinion attribution. Currently, this element is either included in the content span or marked as supplement, depending on how it is expressed. Another future addition should include the proposed features. Since different areas of study address different types of ARs, the attribution type would be a relevant aspect to add since it would allow to just select assertions or opinions. Moreover, it would be useful for factuality studies since the attribution type expresses the source's commitment towards the truth of the content and thus has implications on its factuality.

While the automatic detection of ARs is crucial for a number of other studies and applications, most of these tasks would require the extraction to have large coverage and precision. Otherwise, if the risk of missing relevant information or assigning statements to the incorrect source is too high, manual assessment would still be needed. In this light, the 71% strict *F-score* the proposed model achieves is just a starting point. While this might seem not a satisfactory achievement, nor a big step forward with respect to existing approaches, we have to consider the broader scope of attribution in the present work. This result is obtained for the extraction of a wide range of ARs, of which only subsets were previously identified. Identified attributions include:

- Direct, indirect and mixed ARs.
- Quotations, opinions and other types of ARs.
- The separate identification of source, cue and content spans.
- Inter-sentential ARs.
- A broad range of sources (e.g. nominal, pronominal), cues (e.g. verbal, nominal, adjectival) and content spans (e.g. clausal, non-clausal).

Nonetheless, several challenges remain to be addressed. The main limitations of the current approach are that it cannot identify nested ARs and it does fail to recognize whether an AR has an implicit source. Moreover, while the majority of attributional cues can be reliably identified by the cue classifier, the system does not handle cues other than verbs, with the exception of the prepositional group 'according to' and a small hand-built list of noun-cues. The system is mostly hindered by the relatively low strict recall of the content extraction component. This is affected by the presence of nested AR cues and by inter-sentential and attachment ambiguity.

The attribution extraction model paves the way for the development of attribution-based applications and further studies on attribution. I have presented preliminary investigations of the applicability of the current approach to other genres and languages in Ch. 5 and explored the possibility to apply attribution to news summarization and to information extraction in Ch. 6.

Appendix A

Inter-annotator Agreement Study - Annotation Schema and Instructions

A.1 Attribution Identification

Attribution is a relation identifying a third party as the owner of an attitude towards some text. This can be an utterance, a belief or knowledge or an intention. An attribution is typically composed by three elements (see Ex.(135)): the source, the **cue** and the *content*.

(135) *"The morbidity rate is a striking finding among those of us who study asbestos-related diseases,"* **said** Dr. Talcott.

Four types of attribution are annotated:

- assertions (Ex.(136))(say, write, smile, ...), i.e. acts of communication, even if implicit, e.g with manner verbs (smile —>said while smiling)
- beliefs (idea, think, believe ...), i.e. the expression of a mental process
- facts (know, see, hear, ...), i.e. when the content is presented as a fact
- eventualities (INFLUENCE: order, appoint; COMMITMENT: agree, promise, accept; ORIENTATION: hope, want, ...)

Emotions (e.g. John is happy) are not in the scope of the annotation. Idiomatic attributions, e.g. *it is to say*, should not be annotated. The content should express the attributed linguistic material (Ex. 136a) or its semantic content (Ex. 136b) and not just an **empty attribution**, i.e. a description of what was expressed (Ex. 136c).

(136) a. V John **said**: *"I am sorry"*.

b. V John **said** *that he is sorry*.

c. X John said three words.

In case the empty attribution is anaphorically referring to the actual content, expressed somewhere else in the article, this should be annotated (Ex. (137)).

(137) V ”I am sorry”. John **said** *these three words*.

A.2 Annotation with MMAX2

The annotation process starts with loading an article at a time in the MMAX2 tool and is done at once for each attribution relation, by performing four different steps.

N.B.: Please remember to always make sure Settings>Auto-apply is selected (otherwise your attribute selection will not be saved) and to save the annotation when closing the article or loading a new one.

1. select each markable (i.e. span to annotate) part of the attribution relation (they get displayed in blue bold text in between square brackets) (Sec. A.2.1)

2. assign a role to each markable, i.e. source/cue/content/supplement (each identified by a different colour background) (Sec. A.2.1)

3. select the cue markable and assign values to each attribute (Sec. A.2.2)

4. link all the markables in a relation: right-click on the markable to include in the set (linked markables are displayed joined by red arches) (Sec. A.2.3)

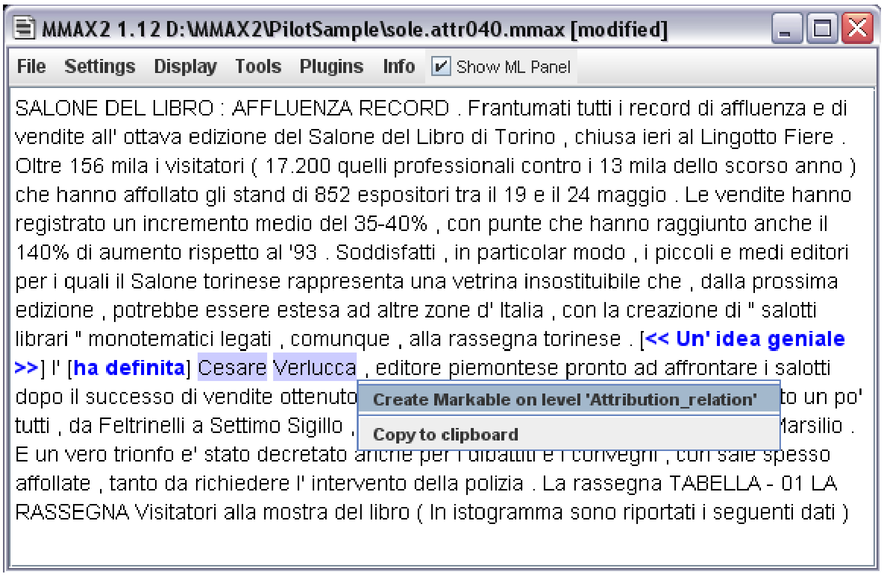
A.2.1 Markables Selection and Labelling

After having identified an attribution relation, the relevant text spans need to be selected and labelled as markables. Please NOTE that an attribution can occur inside the content of another attribution (e.g. [John] [**thinks**] [[*Mary*] [***believes***] [*that …*]]). Once an attribution relation is found, it is necessary first of all to identify its constitutive elements **source**, **cue**, **content** and **supplement** and determine which span represents them. Each relation has at least three components:

- the **cue**, i.e. the textual anchor signalling the relation

- the *content*, i.e. the attributed material
- the source, i.e. the entity the content is attributed to (possibly implicit)

In some cases it is instead possible to have multiple instances of ‘source’(e.g. consultants and industry executives said) and ‘content’(e.g. The president **said** [*that the economy is on the verge of a severe crisis*] *and* [*that he is going to meet the ministers to talk about possible solutions*].). In addition to these three components there is a fourth one, the SUPPLEMENT, which can be optionally used to mark additional relevant information.


MMAX2 1.12 D:\MMAX2\PilotSample\sole.attr040.mmax [modified]
File
Settings
Display
Tools
Plugins
Info
☑ Show ML Panel
Create Markable on level 'Attribution_relation'
Copy to clipboard

SALONE DEL LIBRO : AFFLUENZA RECORD . Frantumati tutti i record di affluenza e di vendite all' ottava edizione del Salone del Libro di Torino , chiusa ieri al Lingotto Fiere . Oltre 156 mila i visitatori (17.200 quelli professionali contro i 13 mila dello scorso anno) che hanno affollato gli stand di 852 espositori tra il 19 e il 24 maggio . Le vendite hanno registrato un incremento medio del 35-40% , con punte che hanno raggiunto anche il 140% di aumento rispetto al '93 . Soddisfatti , in particolar modo , i piccoli e medi editori per i quali il Salone torinese rappresenta una vetrina insostituibile che , dalla prossima edizione , potrebbe essere estesa ad altre zone d' Italia , con la creazione di " salotti librari " monotematici legati , comunque , alla rassegna torinese . [<< Un' idea geniale >>] l' [ha definita] Cesare Verlucca , editore piemontese pronto ad affrontare i salotti dopo il successo di vendite ottenuto to un po' tutti , da Feltrinelli a Settimo Sigillo , larsilio . E un vero trionfo e' stato decretato anche per i dibattiti e i convegni , con sale spesso affollate , tanto da richiedere l' intervento della polizia . La rassegna TABELLA - 01 LA RASSEGNA Visitatori alla mostra del libro (In istogramma sono riportati i seguenti dati)

Figure A.1: Annotation, text spans selection

The text spans corresponding to cue, source and content should be first selected (as in Table A.1) thus enabling the option of creating a markable with the selected text. In case extensions or reductions to the text span corresponding to a markable are required, it is possible to do so with choosing ‘add’or ‘remove from this markable’from the menu on the selected span. Elements that can possibly constitute each markable type are listed in Figure A.2 (cues can also be expressed by adverbials, e.g. allegedly, reportedly).

Deciding what is in the scope of the attribution relation, i.e. what exactly to comprise in each markable, should not be taken for granted. In the following sections indications will be provided about each markable type and what should be included or left out of its text span.

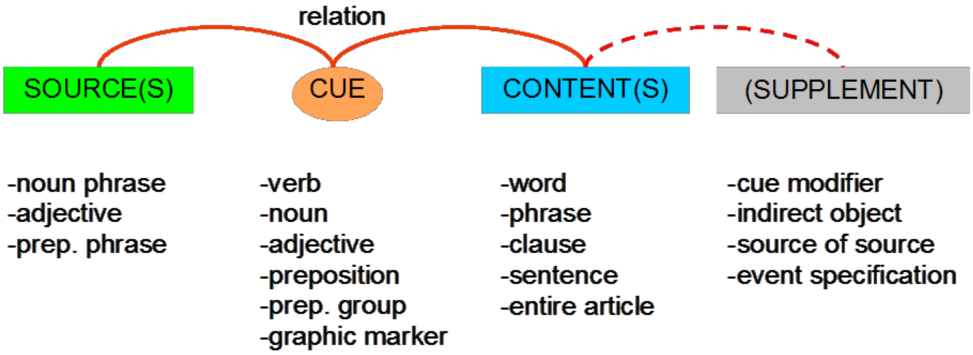


Figure A.2: Markables elements

A.2.1.1 Source Span

In general, in the source span should be included all those elements relevant to the identification of the entity having this role. The entity can be named (Mr. Smith) or unnamed (e.g. a man), animate or inanimate (e.g. The White House/ the article) or even implicit (e.g. It was reported that...). The source markable should always comprehend the full noun phrase expressing it. In case of appositives or relative clauses referring to the entity in the noun phrase and contributing to its characterisation, these should also be selected together with the noun phrase (Ex.(138a) and (138b)). In case the source is represented by an adjective (e.g. the presidential report) or a possessive pronoun (Ex.(139)), the full noun group should be annotated. Implicit sources do not have a corresponding markable since they are not expressed in the text. Null or missing subjects, having no corresponding span, should also not be marked.

(138) a. "...", said Sterling Pratt, wine director at Schaefer's in Skokie, Ill., one of the top stores in suburban Chicago .

b. ... says Warren H. Strother, a university official who is researching a book on Mr. Hahn.

(139) His advice:"Don't panic".

When the relation is part of a relative clause with the source expressed by a relative pronoun, just the pronoun should be annotated as in Ex. (140). Finally, attribution should not be confused with evidence. Compare Ex.(141a) and (141b). In the second example the pseudo-source is just the evidence allowing the writer to draw the conclusion (pseudo-content).

(140) Bay Financial, which said it may be forced to file under Chapter 11 if it can't reach an agreement with its lenders to relieve its debt burden, plunged 1 3/8 to 2 1/8.

(141) a. The report **shows** *that deaths on urban interstate highways rose 7% between 1986 and ...*

b. The figures from the past few years show that deaths on urban interstate highways rose 7% between 1986 and ...

A.2.1.2 Cue Span

The cue can be expressed by a considerable number of elements thus making it difficult to automatically recognise it. Most commonly, however, cues are verbs, not only reporting verbs (e.g. say, write, confirm, think), but also manner verbs (e.g. shrug, beam) and other verbs (e.g. add, continue). Verbal cues should be annotated together with their full verbal group, including auxiliaries, modals and negative particles (e.g. he **didn't say**). Adverbials adjacent to the cue (e.g. she **said angrily**) need to be included, since they can modify the verb. Other elements part of the verbal phrase can be marked as supplement (e.g. she **said** WITH ANGER).

Occasionally cues are expressed by other elements as listed in Figure A.2. Relatively frequent are: prepositions or prepositional groups (e.g. according to, for, in the eyes of) and nouns (e.g. report, idea, fear) as in (142). While cues of different types should be split into separate attribution relations, those of the same type concur to signalling the presence of an attribution and should be grouped. An exception is made only for punctuation cues which should be annotated only when the relation is not signalled by other means as in Ex. (143)).

(142) a. There is **evidence** *that if people inherit defective versions of these genes ...*

b. Our **hope** *that the Senator and other members of the congressional left...*

c. *Even the volatility created by stock index arbitrage and other computer-driven trading strategies isn't entirely bad*, in Mr. Connolly**'s view** .

(143) Rep. Mary Rose Oakar (D., Ohio) at last week's hearings on irregularities in programs at the Department of Housing and Urban Development**:** I don't want to feel guilty representing my constituents ...

A.2.1.3 Content Span

The selection of the content should obey to a principle of limiting the annotation to that portion of text which is confidently perceived as meant to be attributed to the source. This means that the content span should not include utterances of uncertain attribution due to syntactic ambiguities. An example is when a clause constituting the content is joined to another utterance via a coordinating conjunction. In this case, only if the complementizer *that* is included (Ex. (144)) the second clause is also surely attributed, otherwise it could represent material added by the writer.

(144) Still, without many actual deals to show off, Kidder is left to stress *that it finally has "a team" in place*, and *that everyone works harder*.

When the content span is separated by an incidental phrase or clause, it should be annotated as a single markable, unless, as Ex.(145), the content is also divided by sentence boundaries. In this case it seems more appropriate the addition of the second part of the attribution still to the same relation, though as a second content markable.

(145) (154) *"There's no question that some of those workers and managers contracted asbestos-related diseases,"* **said** Darrell Phillips, vice president of human resources for Hollingsworth & Vose. *"But you have to recognize that these events took place 35 years ago. It has no bearing on our work force today."*

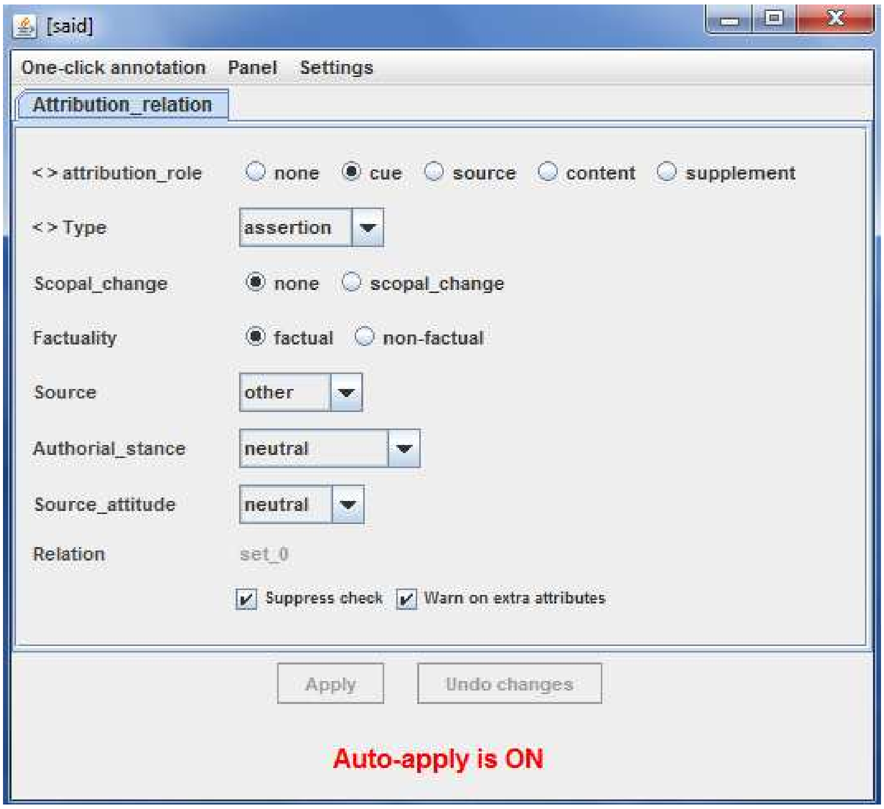
The complementizer THAT should always be included in the content span, together with the QUOTATION MARKS. Punctuation at the end of a content span should only be included if part of the content itself. This means that for example a full stop at the end should be included when the content is expressed by a full sentence, a question mark when the content itself is a question and so forth (or when inside the quotation marks).

A.2.1.4 Supplement Span

As supplement are annotated additional elements which, although not fundamental in an attribution relation, do carry useful information. These can be: concurring to the identification of the source and the provenance or mean by which the information was acquired (e.g. said ON THE PHONE); providing further specification of the attitude this holds (e.g. said WITH ANGER); the recipient of a reportive verb of the assertion type (e.g. told THE JURY) or of an eventuality (e.g. Mary expects JOHN to do the shopping); and event specifications (e.g. said LAST WEEK) providing context indications determinant to the interpretation and comprehension of the content.

A.2.2 Feature Annotation Guidelines

After selecting the text spans corresponding to the elements part of an attribution relation it is necessary to assign the role to each markable in the 'annotation window'(Figure A.3). When the role 'cue'is chosen, the window will display also the attributes and their values which need to be assigned. The feature 'scopal change'is disabled when cues of the type 'fact'are selected.



<> attribution_role ○ none ● cue ○ source ○ content ○ supplement

<> Type: assertion

Scopal_change ● none ○ scopal_change

Factuality ● factual ○ non-factual

Source: other

Authorial_stance: neutral

Source_attitude: neutral

Relation: set_0

☑ Suppress check ☑ Warn on extra attributes

Figure A.3: Attribute selection

A.2.2.1 Type Attribute

The type of attitude held by the source is by default NONE. In the annotation window however, one of the four values this feature can assume, namely ASSERTION, BELIEF, FACT and EVENTUALITY, needs to be selected. The preposition **for** and **according to** are considered assertions. Other prepositional groups (e.g. in the opinion of, in the eyes of, in the perspective of) should be marked as 'belief'. Verb cues need instead to be considered in context and annotated according to the attitude they express. For example, 'ask'is an assertion in Ex.(146a) and an eventuality in Ex.(146b).

(146) a. John **asked** MARY *"are you happy?"*.

b. John **asked** MARY *to be ready by 7pm.*

A.2.2.2 Factuality Attribute

The factuality attribute takes just two values: FACTUAL and NON-FACTUAL. In order to decide which value to assign, it is necessary to concentrate on the attribution relation itself no matter what the content is. 'Factual'is by default the value assigned, it is in fact more frequent, at least in journalistic texts, and represents real attributions. In case the attribution relation is not a real bound but just an hypothetical match (Ex.(147)) or the negation of a link between source and content (Ex.(148)), it takes the value 'non-factual'.

(147) Network officials involved in the studio talks **may hope** *the foreign influx builds more support in Washington.*

(148) Mr. Smith **didn't say** *that he will take the a part in the film.*

The factuality can be compromised by the following elements when they scope on the cue:

- polarity reversing particle (negation, negative pronouns)
- verb mode (conditional, imperative)
- verb tense (future)
- hypothetical (if)
- interrogative form
- modals (could, might, would, . . .)

The factuality judgement represents the answer to the following question: **is the attribution of the content to the source presented as a fact in the real world?**

A.2.2.3 Scopal Change Attribute

Also the scopal change attribute can take two values, NONE being the default one, and SCOPAL CHANGE. A change in the scope happens relatively seldom, however it is important to recognise it in order to avoid incorrectly considering it as affecting the factuality. The scopal change almost solely occurs with polarity, therefore it is opportune to pay particular attention to those attributions appearing at first as non-factual because of the cue being in the scope of a negation, or entailing negation (e.g. deny Ex.(149)). In these cases it can be checked if there is a polarity change first with determining whether there is still a perceived attribution and secondly with considering if the reverse of the content is attributed, i.e. the negation could be moved to the content (e.g from Ex.(149): He furthermore **says** that he DIDN'T rely too heavily on . . .).

(149) He furthermore **denies** that he relied too heavily on Sotheby's or Mr. Wachter.

In case of eventualities or beliefs, the scopal change refers to the fact that it is not the polarity of the attribution that is affected (the attribution is factual) nor that of the content, but rather the polarity of the attitude held by the source (e.g. 'John **doesn't want** us to take a holiday'means that **not wanting** is the attitude, which is not necessarily the same as 'John wants us not to take a holiday').

A.2.2.4 Source Type Attribute

The source is by default WRITER and can assume also the values: OTHER, ARBITRARY and MIXED. 'Writer'should be assigned in case the attribution is overtly to the writer of the article (usually expressed by I) while 'other'refers to another defined entity, including very general sources like a man or experts. As 'arbitrary'should be marked those instances without a specific source, i.e. impersonal or hidden sources such as everyone, the people, one or pronouns like you or they when used as impersonals. 'Mixed'should be instead used to mark when an attribution possesses multiple sources of different type (e.g. The president and everyone think).

A.2.2.5 Authorial Stance

This features marks the commitment of the author towards the truth of what is expressed by the content. The author is COMMITTED, if the content is presented as truthful. This is usually the case with cue verbs like: *admit*, *confess*, *acknowledge*, *know*, *recognize*, *realize* and in general with attribution of the type *fact*. On the contrary, if the author is suggesting that what is expressed in the content is not truthful, e.g with cue verbs like lie or joke, this should be marked as NON-COMMITTED. Most verbs, e.g. *say*, *suggest*, *believe*, *suspect*, *deny*, express a rather NEUTRAL stance (default value).

A.2.2.6 Source Attitude

The source attitude marks the attitude the source expresses towards the content. This can be POSITIVE (e.g. welcome, marvel, congratulate), CRITICAL (e.g. fear, protest, lament), TENTATIVE (e.g. think, believe, suggest), NEUTRAL (default value) (e.g. say, comment, add) or OTHER. The attitude is usually identified by the choice of attributional verb, in particular manner verbs carry an attitude (e.g. smile (positive), sniff (critical)). Among other contextual elements that can also express the attitude: source modifiers (e.g. a smiling Mr. Smith said); prepositional phrases (e.g. said in an uncertain voice); adverbials (e.g. said angrily).

A.2.3 Markables Linking

The last required step is that of linking together all the markables in an attribution relation. This is done by selecting one of the markables and then right-cliking on each markable to add to the set and selecting the appropriate option from the menu. Linked markables are displayed connected by red arches.

Appendix B

PARC 3.0 - Annotation Schema and Instructions

B.1 Attribution Identification

Attribution is a relation identifying a third party as the owner of an attitude towards some text. This can be an utterance, a belief or knowledge or an intention. An attribution is typically composed by three elements (see Ex.(150)): the source, the cue and the content.

(150) "The morbidity rate is a striking finding among those of us who study asbestos-related diseases," said Dr. Talcott

Four types of attribution are annotated:

- assertions (Ex.(151))(say, write, smile, ...), i.e. acts of communication, even if implicit, e.g with manner verbs (smile —>said while smiling)
- beliefs (idea, think, believe ...), i.e. the expression of a mental process
- facts (know, see, hear, ...), i.e. when the content is presented as a fact
- eventualities (INFLUENCE: order, appoint; COMMITMENT: agree, promise, accept; ORIENTATION: hope, want, ...)

Emotions (e.g. John is happy) are not in the scope of the annotation. Idiomatic attributions, e.g. *it is to say*, should not be annotated. The content should express the attributed linguistic material (Ex. 151a) or its semantic content (Ex. 151b) and not just an **empty attribution**, i.e. a description of what was expressed (Ex. 151c).

(151) a. V John said: “I am sorry”.

b. V John said that he is sorry.

c. X John said three words.

In case the empty attribution is anaphorically referring to the actual content, expressed somewhere else in the article, this should be annotated (Ex. (152)).

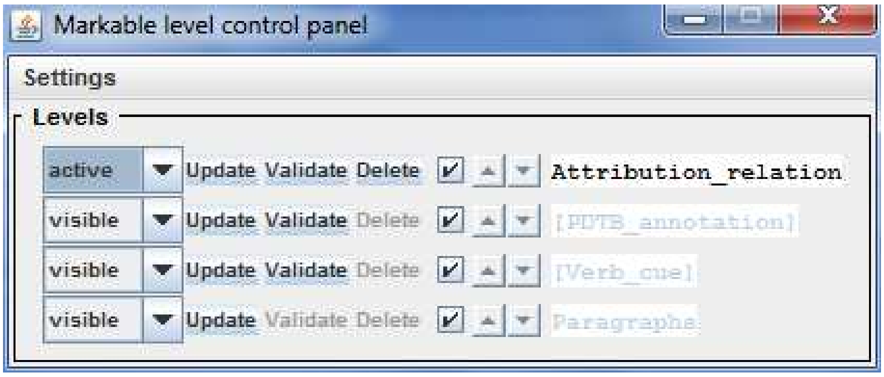
(152) V ”I am sorry”. John said these three words

B.2 Annotation with MMAX2

The annotation process starts with loading an article at a time in the MMAX2 tool and is done at once for each attribution relation by performing a set of steps.

B.2.1 Pre-steps

1. In the ’Markable level control panel’ set the levels: PDTB_annotation, Verb_cue and Paragraphs to VISIBLE, leaving only Attribution_realtion as ACTIVE.



Markable level control panel

Settings

Levels

active ▼ Update Validate Delete ☑ ▲ ▼ Attribution_relation

visible ▼ Update Validate Delete ☑ ▲ ▼ [PDTB_annotation]

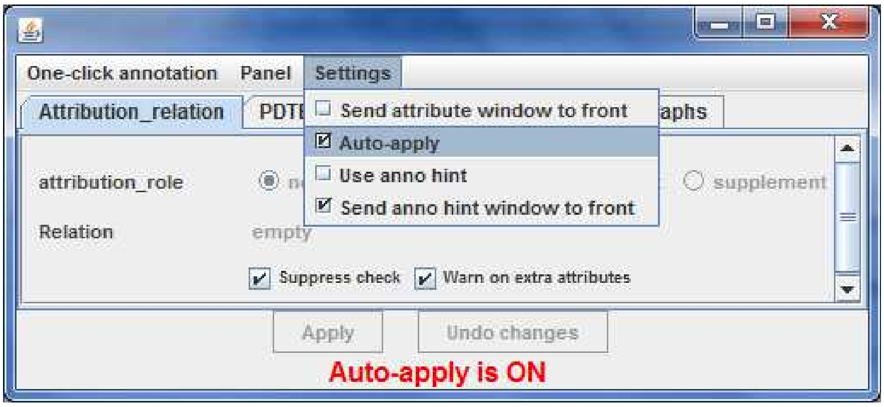
visible ▼ Update Validate Delete ☑ ▲ ▼ [Verb_cue]

visible ▼ Update Validate Delete ☑ ▲ ▼ Paragraphs

2. In the ’Annotation panel’ tick Settings>Auto-apply.

3. In the ’Text panel’ tick Settings>Select new markable after creation.

4. Proceed with the annotation. (If the annotation is not displayed correctly, try pressing F5 on your keyboard – this corresponds to: Display>Reapply current style sheet).



□ Send attribute window to front
☑ Auto-apply
□ Use anno hint
☑ Send anno hint window to front

attribution_role ● n... ○ supplement

Relation empty

☑ Suppress check ☑ Warn on extra attributes

Auto-apply is ON

5. Once finished, you can save the annotations and exit or switch to another file. MMAX2 asks you whether to save or discard the annotations before closing an unsaved file.

N.B.: Please remember to always make sure Settings>Auto-apply is selected (otherwise your attribute selection will not be saved) and to save the annotation when closing the article or loading a new one.

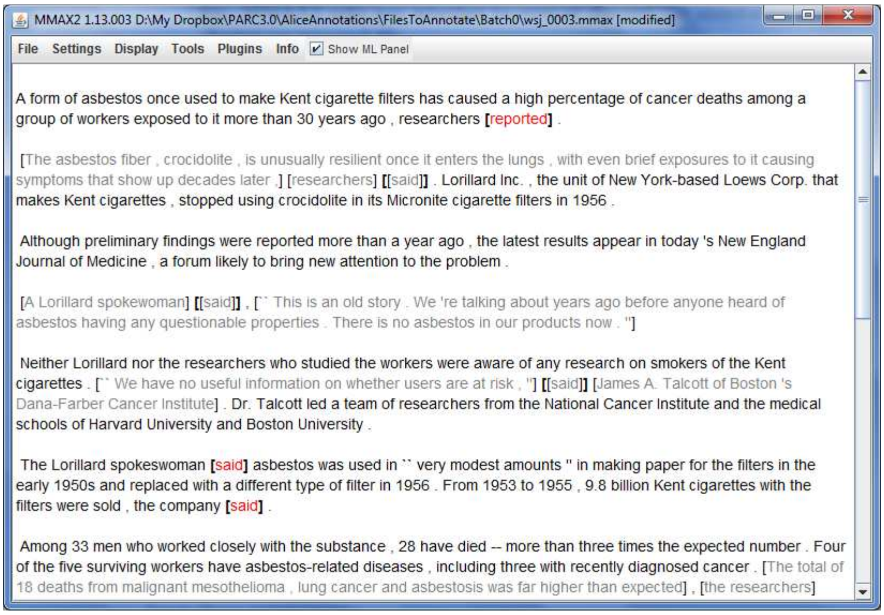
B.2.2 Levels

At this stage you are asked to **annotate only the black text** and **ignore the gray text**, which corresponds to already annotated attribution relations. **Do not look for attributions inside the gray text.** You are asked to annotated the first level of attributions, that is, **do not look for a nested attribution inside an annotation you just made**. If you annotated an attribution and realize that it is nested into another one, just proceed with annotating the outer one (but don't delete the one you already annotated).

Some verbs and 'according to' are marked in red. Those are the verbs that the automatic cue classifier has identified as **potential attribution cues**. They are intended as a support to the annotation. However, the classifier can be wrong. **Don't expect to find an attribution for each marked verb, and expect to find verb cues the classifier did not identify**. The classifier is also unable to handle cues other than verbs (and 'according to').

B.2.3 Annotation steps

1. Select each markable (i.e. span to annotate) part of the attribution relation (Sec. B.3).


MMAX2 1.13.003 D:\My Dropbox\PARC3.0\AliceAnnotations\FilesToAnnotate\Batch0\wsj_0003.mmax [modified]
File
Settings
Display
Tools
Plugins
Info
☑ Show ML Panel

A form of asbestos once used to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago , researchers [reported] .

[The asbestos fiber , crocidolite , is unusually resilient once it enters the lungs , with even brief exposures to it causing symptoms that show up decades later .] [researchers] [[said]] . Lorillard Inc. , the unit of New York-based Loews Corp. that makes Kent cigarettes , stopped using crocidolite in its Micronite cigarette filters in 1956 .

Although preliminary findings were reported more than a year ago , the latest results appear in today 's New England Journal of Medicine , a forum likely to bring new attention to the problem .

[A Lorillard spokewoman] [[said]] , [`` This is an old story . We 're talking about years ago before anyone heard of asbestos having any questionable properties . There is no asbestos in our products now . '']

Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes . [`` We have no useful information on whether users are at risk , ''] [[said]] [James A. Talcott of Boston 's Dana-Farber Cancer Institute] . Dr. Talcott led a team of researchers from the National Cancer Institute and the medical schools of Harvard University and Boston University .

The Lorillard spokeswoman [said] asbestos was used in `` very modest amounts '' in making paper for the filters in the early 1950s and replaced with a different type of filter in 1956 . From 1953 to 1955 , 9.8 billion Kent cigarettes with the filters were sold , the company [said] .

Among 33 men who worked closely with the substance , 28 have died -- more than three times the expected number . Four of the five surviving workers have asbestos-related diseases , including three with recently diagnosed cancer . [The total of 18 deaths from malignant mesothelioma , lung cancer and asbestosis was far higher than expected] , [the researchers]

2. Assign a role to each markable, i.e. source/cue/content/supplement (each identified by a different colour background) (Sec. B.3).

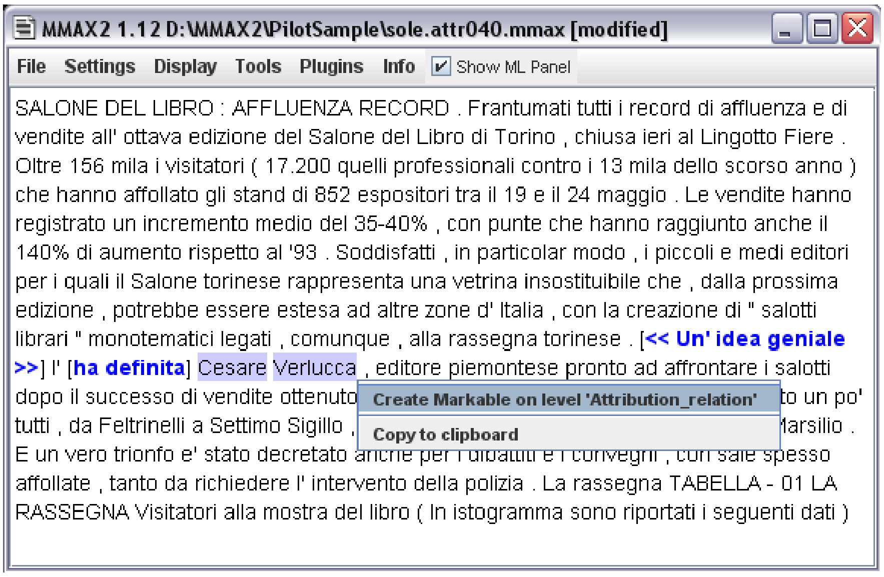
3. Link all the markables in a relation: right-click on the markable to include in the set (linked markables are displayed joined by red arches) (Sec. B.4).

B.3 Markables Selection and Labelling

After having identified an attribution relation, the relevant text spans need to be selected and labelled as markables. Once an attribution relation is found, it is necessary first of all to identify its constitutive elements source, cue, content and the optional supplement and determine which span represents them. Each relation has at least three components:

- the cue, i.e. the textual anchor signalling the relation
- the content, i.e. the attributed material
- the source, i.e. the entity the content is attributed to (possibly implicit)

In some cases it is instead possible to have multiple instances of 'source'(e.g. consultants and industry executives said) and 'content'(e.g. The president said [that the economy is on the verge of a severe crisis] and, addressing the Parliament, [that he is going to meet the ministers to talk about possible solutions].). In addition to these three components there is a fourth one, the supplement, which can be optionally used to mark additional relevant information


MMAX2 1.12 D:\MMAX2\PilotSample\sole.attr040.mmax [modified]
File
Settings
Display
Tools
Plugins
Info
☑ Show ML Panel
Create Markable on level 'Attribution_relation'
Copy to clipboard

SALONE DEL LIBRO : AFFLUENZA RECORD . Frantumati tutti i record di affluenza e di vendite all' ottava edizione del Salone del Libro di Torino , chiusa ieri al Lingotto Fiere . Oltre 156 mila i visitatori (17.200 quelli professionali contro i 13 mila dello scorso anno) che hanno affollato gli stand di 852 espositori tra il 19 e il 24 maggio . Le vendite hanno registrato un incremento medio del 35-40% , con punte che hanno raggiunto anche il 140% di aumento rispetto al '93 . Soddisfatti , in particolar modo , i piccoli e medi editori per i quali il Salone torinese rappresenta una vetrina insostituibile che , dalla prossima edizione , potrebbe essere estesa ad altre zone d' Italia , con la creazione di " salotti librari " monotematici legati , comunque , alla rassegna torinese . [<< Un' idea geniale >>] l' [ha definita] Cesare Verlucca , editore piemontese pronto ad affrontare i salotti dopo il successo di vendite ottenuto to un po' tutti , da Feltrinelli a Settimo Sigillo , Marsilio . E un vero trionfo e' stato decretato anche per i dibattiti e i convegni , con sale spesso affollate , tanto da richiedere l' intervento della polizia . La rassegna TABELLA - 01 LA RASSEGNA Visitatori alla mostra del libro (In istogramma sono riportati i seguenti dati)

Figure B.1: Annotation, text spans selection

The text spans corresponding to cue, source and content should be first selected (as in Table B.1) thus enabling the option of creating a markable with the selected text. In case extensions or reductions to the text span corresponding to a markable are required, it is possible to do so with choosing **'add'or 'remove from this markable'** from the menu on the selected span. Elements that can possibly constitute each markable type are listed in Figure B.2 (cues can also be expressed by adverbials, e.g. allegedly, reportedly).

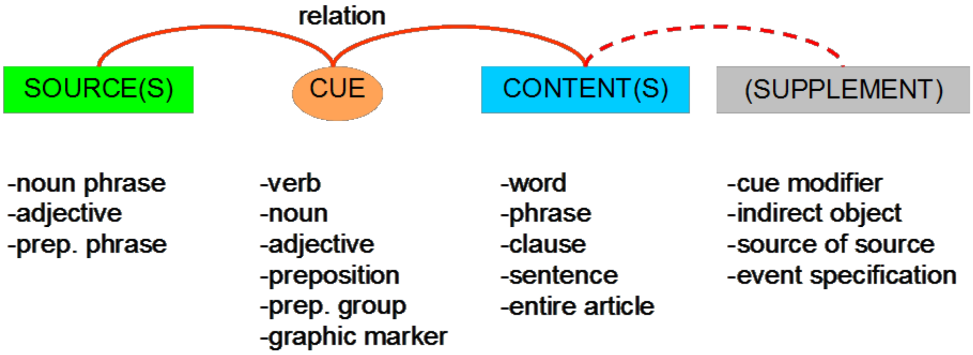


Figure B.2: Markables elements.(Old colour scheme: the SOURCE will be RED and the CONTENT ORANGE. GRAY will highlight non-classified markables.

Deciding what is in the scope of the attribution relation, i.e. what exactly to comprise in each markable, should not be taken for granted. In the following sections indications will be provided about each markable type and what should be included or left out of its text span.

B.3.1 Source Span

In general, the source span should **include all those elements relevant to the identification of the entity** having this role. The entity can be named (Mr. Smith) or unnamed (e.g. a man), animate or inanimate (e.g. The White House/ the article) or even implicit (e.g. It was reported that...). The source markable should always comprehend the **full noun phrase** expressing it. In case of **appositives** or **relative clauses** referring to the entity in the noun phrase and contributing to its characterisation, these should also be selected together with the noun phrase (Ex.(153a) and (153b)). In case the source is represented by an adjective (e.g. the presidential report) or a possessive pronoun (Ex.(154)), the full noun group should be annotated. Implicit sources do not have a corresponding markable since they are not expressed in the text. Null or missing subjects, having no corresponding span, should also not be marked.

(153) a. "...", said Sterling Pratt, wine director at Schaefer's in Skokie, Ill., one of the top stores in suburban Chicago .

b. ...says Warren H. Strother, a university official who is researching a book on Mr. Hahn.

(154) His advice:"Don't panic".

When the relation is part of a relative clause with the source expressed by a **relative pronoun**, just the pronoun should be annotated as source, as in Ex. (155).

(155) Bay Financial, which said it may be forced to file under Chapter 11 if it can't reach an agreement with its lenders to relieve its debt burden, plunged 1 3/8 to 2 1/8.

Finally, attribution should not be confused with evidence. Compare Ex.(156a) and (156b). In the second example the pseudo-source is just the evidence allowing the writer to draw the conclusion (pseudo-content).

(156) a. The report shows that deaths on urban interstate highways rose 7% between 1986 and . . .

b. The figures from the past few years show that deaths on urban interstate highways rose 7% between 1986 and . . .

B.3.2 Cue Span

The cue can be expressed by a considerable number of elements thus making it difficult to automatically recognise it. Most commonly, however, cues are verbs, not only reporting verbs (e.g. say, write, confirm, think), but also manner verbs (e.g. shrug, beam) and other verbs (e.g. add, continue). Verbal cues should be annotated together with their **full verbal group**, including auxiliaries, modals and negative particles (e.g. he didn't say). Adverbials adjacent to the cue (e.g. she said angrily) need to be included, since they can modify the verb. Other elements part of the verbal phrase can be marked as supplement (e.g. she said with anger).

Occasionally cues are expressed by other elements as listed in Figure B.2. Relatively frequent are: prepositions or prepositional groups (e.g. according to, for, in the eyes of) and nouns (e.g. report, idea, fear) as in (157). While cues of different types should be split into separate attribution relations (even if that means they share the same source and/or content, e.g. he says and believes), those of the same type concur to signalling the presence of an attribution and should be grouped. An exception is made only for punctuation cues which should be annotated only when the relation is not signalled by other means as in Ex. (158)).

(157) a. There is evidence that if people inherit defective versions of these genes . . .

b. Our hope that the Senator and other members of the congressional left. . .

c. Even the volatility created by stock index arbitrage and other computer-driven trading strategies isn't entirely bad, in Mr. Connolly's view .

(158) Rep. Mary Rose Oakar (D., Ohio) at last week's hearings on irregularities in programs at the Department of Housing and Urban Development: I don't want to feel guilty representing my constituents . . .

B.3.3 Content Span

The selection of the content should obey to a principle of limiting the annotation to that portion of text which is confidently perceived as meant to be attributed to the source. This means that the content span should not include utterances of uncertain attribution due to syntactic ambiguities. An example is when a clause constituting the content is joined to another utterance via a coordinating conjunction. In this case, only if the complementizer *that* is included (Ex. (159)) the second clause is also surely attributed, otherwise it could represent material added by the writer.

(159) Still, without many actual deals to show off, Kidder is left to stress that it finally has ”a team” in place, and that everyone works harder.

When the content span is separated by an incidental phrase or clause, it should be annotated as a single markable, unless, as Ex.(160), the content is also divided by sentence boundaries. In this case it seems more appropriate the addition of the second part of the attribution still to the same relation, though as a second content markable.

(160) (154) ”There's no question that some of those workers and managers contracted asbestos-related diseases,” said Darrell Phillips, vice president of human resources for Hollingsworth & Vose. ”But you have to recognize that these events took place 35 years ago. It has no bearing on our work force today.”

The complementizer THAT should always be included in the content span, together with the QUOTATION MARKS. Punctuation at the end of a content span should only be included if part of the content itself. This means that for example a full stop at the end should be included when the content is expressed by a full sentence, a question mark when the content itself is a question and so forth (or when inside the quotation marks).

B.3.4 Supplement Span

As supplement are annotated additional elements which, although not fundamental in an attribution relation, do carry useful information. These can be: concurring to the identification of the source and the provenance or mean by which the information was acquired (e.g. said on the phone); providing further specification of the attitude this holds (e.g. said with anger); the recipient of a reportive verb of the assertion type (e.g. told the jury) or of an eventuality (e.g. Mary expects John to do the shopping); and event specifications (e.g. said last week) providing context indications determinant to the interpretation and comprehension of the content.

B.4 Markables Linking

The last required step is that of linking together all the markables in an attribution relation. This is done by **selecting one of the markables and then right-cliking on each markable to add it to the set** and selecting the appropriate option from the menu. Linked markables are displayed connected by red arches.

B.5 Doubts, Solutions and Harder Cases

B.5.1 Sets Sharing Elements

Attribution relations can have 0-N sources and supplements and 1-N content spans. But, ONE CUE = ONE ATTRIBUTION RELATION.

(161) Newsweek, trying to keep pace with rival Time magazine, announced new advertising rates for 1990 and said it will introduce a new incentive plan...

Solution:

(162)AR 1 Newsweek, trying to keep pace with rival Time magazine, announced new advertising rates for 1990 and said it will introduce a new incentive plan...

AR 2 Newsweek, trying to keep pace with rival Time magazine, announced new advertising rates for 1990 and said it will introduce a new incentive plan...

B.5.2 Errors in the Gray ARs

Do not look at the gray text. You will be asked to correct those ARs in the second annotation stage.

HOWEVER, cases like (163) are not errors. The AR in gray is complete, but nested into another one. The gray text is the content span of the AR 'Mrs. Hills' 'said'. Just annotate the AR as usual.

(163) [Saudi Arabia], for its part, [has vowed] [to enact a copyright law compatible with international standards and to apply the law to computer software as well as to literary works], Mrs. Hills said.

B.5.3 Passive, Negation and Verb Group

Keep the verb group together. The verb-cue classifier only identifies the head of verb group. In the cue markable you are asked to include the complete verb group and eventual modifiers, e.g.:

- was probably announced
- has been repeatedly said
- shouldn't approve

B.5.4 Laws, Courts and Orders

We do annotate laws, sentences and orders.

(164) a. In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos.

b. The Parliament approved a ban on all uses of asbestos.

c. The Court found him guilty.

B.5.5 Recipient

We annotate the recipient/destinatary as: Orders/speech acts: SUPPLEMENT

(165) She told/ordered John to stop following her.

Expectations/opinions: CONTENT, unless in passive form (just for consitency with previous annotation).

(166) a. They expect the Senate to reach an agreement by Monday.

b. The Senate is expected to reach an agreement by Monday.

B.5.6 Possessives Sources

Annotate as source the possessive as well as the possessed entity whether coreferential with the content (e.g. advice/idea/promise) (167a) or itself a source of the content (e.g. book/notes/document) (167c). If the cue-noun is the only cue, it is also annotated as cue (167a,b,d), otherwise not (167c).

(167) a. His advice: His advice — advice

b. Their assumption is: Their assumption — assumption is

c. Her notes recall: Her notes — recall

d. Mr Smith's promise: Mr Smith's promise — 's promise

B.5.7 Attempt, Seek and Try

'Try' is almost always an action and as such it should not be annotated. Actions require some intention to perform them, but this is a step away from how the information is presented in the article (actions and not utterances/intentions/opinions/knowledge).

'Seek' and 'attempt' may be annotated depending on the context. Identify the content and ask yourself if it is meant to expresses an intention of the source (168b) or just report an action (168a). If you are not strongly oriented for an intention, or you need to go one step back to find one, just do annotate these.

(168) a. She's SEEKING clues to the crime =LOOKING FOR >NO

b. She only SEEKS fame and fortune =WANTS >YES

Appendix C

List of Potential Noun Cues

accord
according
accusation
acknowledgement
ad
admission
advice
agreement
allegation
amendment
announcement
answer
anticipation
argument
article
assertion
assumption
assurance
belief
bet
book
bill
call
challenge
charge
chart
citation
claim
command
comment
commercial
complaint
concern
concession
conclusion
condition
confession
confidence
confirmation
consideration
contention
convinction
counterclaim
criticism
critic
cry
data
decision
declaration
deduction
defence
definition
deliberation
demand
denial
depiction
description
dictate
disappointment
disapproval
disclosure
discovery
dispute
document
doubt
effort
elaboration
encouragement
eruption
estimate
eulogy
evidence
exclamation
expectation
explanation
expression
fear
feeling
file
filing
find
finding
figure
forecast
formulation
guess
highlight
hint
hope
idea
illustration
implication
imposition
indication
information
insinuation
inspiration
instruction
intention
interjection
interpretation
issue
joke
knowledge
lament

laugh
law
lawsuit
lecture
legislation
lesson
letter
list
menace
mention
message
mind
moan
need
news
note
notice
notification
oath
objection
observation
offer
opinion
order
pact
paper
permission
plan
pledge
point
policy
poll
praise
prediction
press
proclamation
project
promise
proposal
protest
prove
provision
question
quotation
realization
reason
recognition
recollection
recommendation
recount
reflection
reform
refusal
rejection
remark
repetition
reply
report
reproach
request
requirement
research
resentment
response
revelation
rule
rumor
saying
scream
shout
sigh
sign
signal
snort
specification
speculation
spell
statement
statistic
story
strategy
study
suggestion
suit
support
supposition
survey
suspicion
talk
temptation
testimony
theory
thought
threat
understandment
urge
view
voice
want
warning
wisdom
worry
yell

Appendix D

PARC 3.0 Attributional Verbs

Complete list of all 527 verb types used as attributional in PARC 3.0. Verbs are listed in their base form and ordered by the number of times they occurred as the cue of an AR.

Verb	Occurrence	Verb	Occurrence	Verb	Occurrence
say	9017	predict	101	accuse	50
expect	671	cite	95	disclose	50
add	372	see	95	decline	48
think	333	find	93	explain	47
report	313	suggest	91	acknowledge	46
believe	267	claim	84	attribute	46
want	253	contend	79	concede	45
note	241	show	79	have	45
agree	233	indicate	76	urge	45
tell	191	post	76	admit	44
announce	186	decide	66	recall	44
plan	175	insist	66	allege	41
hope	136	declare	58	charge	41
consider	131	propose	57	offer	40
estimate	130	warn	56	conclude	39
know	128	complain	55	write	38
ask	127	require	54	worry	37
call	122	deny	51	fear	36
argue	121	intend	51	feel	35

Verb	Occurrence	Verb	Occurrence	Verb	Occurrence
confirm	34	project	18	hint	11
describe	34	realize	18	hold	11
promise	34	vow	18	prefer	11
rule	33	contribute	17	reiterate	11
assume	32	express	17	saw	11
figure	32	look	17	accept	10
order	32	plead	17	bet	10
seek	32	doubt	16	consent	10
refuse	31	forecast	16	mention	10
recommend	30	respond	16	seem	10
view	30	value	16	continue	9
allow	28	emphasize	15	convict	9
assert	28	favor	15	imply	9
approve	27	persuade	15	mean	9
comment	26	put	15	praise	9
caution	24	reply	15	quote	9
demand	24	talk	15	refer	9
oppose	24	criticize	14	wish	9
speculate	24	discover	14	authorize	8
advise	23	recognize	14	defend	8
question	23	request	14	define	8
like	22	support	14	felt	8
maintain	22	suppose	14	inform	8
anticipate	21	threaten	14	make	8
blame	21	unveil	14	pledge	8
concern	20	learn	13	point	8
discuss	20	prohibit	13	portray	8
stress	20	reject	13	read	8
wonder	20	signal	13	regard	8
observe	19	determine	12	force	7
state	19	encourage	12	give	7
suspect	19	provide	12	guarantee	7
understand	19	reveal	12	illustrate	7
convince	18	specify	12	invite	7

Verb	Occurrence	Verb	Occurrence	Verb	Occurrence
remark	7	forbid	4	dictate	3
remind	7	impose	4	disappoint	3
rumor	7	indict	4	discourage	3
testify	7	instruct	4	dub	3
try	7	interest	4	entice	3
assure	6	notice	4	equate	3
confess	6	present	4	hear	3
divide	6	prevent	4	imagine	3
file	6	reaffirm	4	joke	3
foresee	6	recount	4	justify	3
forget	6	remember	4	lament	3
interpret	6	set	4	laud	3
list	6	sing	4	notify	3
name	6	sniff	4	ponder	3
press	6	speak	4	prepare	3
quip	6	study	4	proclaim	3
reason	6	sue	4	profess	3
reckon	6	term	4	promote	3
boast	5	voice	4	prove	3
dismiss	5	wait	4	rely	3
guess	5	address	3	repeat	3
hail	5	applaud	3	satisfy	3
identify	5	appreciate	3	sentence	3
ignore	5	attempt	3	shout	3
include	5	attest	3	solicit	3
outline	5	await	3	stipulate	3
permit	5	block	3	tout	3
push	5	calculate	3	advocate	2
answer	4	challenge	3	affirm	2
characterize	4	compare	3	aim	2
counter	4	comply	3	allude	2
credit	4	condemn	3	appeal	2
deem	4	decry	3	approach	2
disagree	4	denounce	3	aspire	2

Verb	Occurrence	Verb	Occurrence	Verb	Occurrence
assail	2	nickname	2	account	1
back	2	object	2	assess	1
ban	2	paint	2	attack	1
bar	2	perceive	2	avoid	1
brag	2	pinpoint	2	battle	1
celebrate	2	prescribe	2	beam	1
chastise	2	pronounce	2	become	1
choose	2	publish	2	beg	1
clarify	2	purr	2	begin	1
clear	2	raise	2	bemoan	1
commit	2	rat	2	bid	1
confide	2	reassure	2	bill	1
contemplate	2	rebuff	2	brim	1
deride	2	record	2	burble	1
discern	2	regret	2	buttress	1
dispute	2	release	2	buy	1
echo	2	resent	2	capture	1
elaborate	2	restate	2	care	1
empower	2	rethink	2	caricature	1
endorse	2	sense	2	castigate	1
ensure	2	snap	2	chide	1
exclaim	2	snort	2	chuckle	1
explore	2	stand	2	commission	1
fret	2	surprise	2	communicate	1
get	2	survey	2	concentrate	1
go	2	theorize	2	concur	1
highlight	2	underscore	2	conspire	1
implore	2	understate	2	construe	1
introduce	2	uphold	2	contain	1
involve	2	volunteer	2	contest	1
label	2	vote	2	convey	1
laugh	2	welcome	2	couch	1
need	2	absolve	1	counsel	1
negotiate	2	acclaim	1	count	1

Verb	Occurrence	Verb	Occurrence	Verb	Occurrence
croon	1	evince	1	jump	1
crow	1	examine	1	lambast	1
cry	1	exclude	1	lay	1
dare	1	exhort	1	lecture	1
deflect	1	exonerate	1	license	1
delight	1	expound	1	limit	1
deliver	1	fantasize	1	link	1
demonstrate	1	fault	1	lobby	1
demur	1	feud	1	love	1
depict	1	flay	1	mail	1
desire	1	flirt	1	maintain	1
detail	1	focus	1	mandate	1
detect	1	frighten	1	marvel	1
develop	1	fume	1	measure	1
diagnose	1	gauge	1	mind	1
direct	1	gloat	1	misstate	1
disapprove	1	grant	1	moan	1
discipline	1	grip	1	mount	1
disclaim	1	grouse	1	muse	1
disincline	1	growl	1	nod	1
dislike	1	grumble	1	nominate	1
disturb	1	gush	1	obligate	1
downgrade	1	harp	1	opt	1
downplay	1	herald	1	pass	1
draw	1	impress	1	pay	1
dream	1	incline	1	peg	1
embrace	1	incorporate	1	place	1
emerge	1	induce	1	please	1
enable	1	influence	1	poise	1
envisage	1	inquire	1	preach	1
envision	1	insinuate	1	preoccupy	1
erupt	1	interject	1	pressure	1
establish	1	investigate	1	presume	1
evaluate	1	irk	1	pretend	1

Verb	Occurrence
prim	1
produce	1
proffer	1
prompt	1
protest	1
purport	1
quash	1
rave	1
reassert	1
re-emphasize	1
reflect	1
reignite	1
relate	1
relieve	1
rename	1
renew	1
renounce	1
repute	1
resist	1
resolve	1
respect	1
restrain	1
review	1
romance	1
ruminate	1
salute	1
schedule	1
score	1
scream	1
send	1
share	1
shrug	1
sigh	1
sign	1
spell	1
sponsor	1
stagewhispers	1
strive	1
swear	1
take	1
teach	1
tear	1
teem	1
terrify	1
trouble	1
trumpet	1
turn	1
underestimate	1
unleash	1
verify	1
wad	1
whisper	1
witness	1
wrestle	1
yell	1

Bibliography

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, M., Draoulec, A. L., Muller, P., Pery-Woodley, M.-P., Prevot, L., Rebeyrolles, J., Tanguy, L., Vergez-Couret, M., and Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Aha, D. and Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Almeida, M. S. C., Almeida, M. B., and Martins, A. F. T. (2014). A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48, Gothenburg, Sweden. Association for Computational Linguistics.

Alrahabi, M. and Desclés, J.-P. (2008). Automatic annotation of direct reported speech in Arabic and French, according to a semantic map of enunciative modalities. In *Proceedings of the 6th international conference on Advances in Natural Language Processing*, GoTAL '08, pages 40–51, Berlin, Heidelberg. Springer.

Alrahabi, M., Descles, J.-P., and Suh, J. (2010). Direct reported speech in multilingual texts: Automatic annotation and semantic categorization. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS '10).*

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34:555–596.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bergler, S. (1992). *Evidential analysis of reported speech*. PhD thesis, Brandeis University, Waltham, MA, USA.

Bergler, S., Doandes, M., Gerard, C., and Witte, R. (2004). Attributions. In Qu, Y., Shanahan, J., and Wiebe, J., editors, *Exploring Attitude and Affect in Text: Theories and Applications*, Technical Report SS-04-07, pages 16–19, Stanford, California, USA. AAAI Press. Papers from the 2004 AAAI Spring Symposium.

Bernardini, L. and Prodanof, I. (2014). L'integrazione di informazioni contestuali e linguistiche nel riconoscimento automatico dell'ironia. In Roberto Basili, Alessandro Lenci, B. M. e., editor, *First Italian Conference on Computational Linguistics CLiC-it 2014*, Pisa, Italy. Pisa University Press.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 22–24.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. " O'Reilly Media, Inc.".

Bloom, K., Stein, S., and Argamon, S. (2007). Appraisal extraction for news opinion analysis at NTCIR-6. In *of NTCIR-6 Workshop Meeting*, Tokyo, Japan.

Bolden, G. (2004). The quote and beyond: defining boundaries of reported speech in conversational Russian. *Journal of pragmatics*, 36(6):1071–1118.

Bracchi, A. (2014). Attribution relation cues across genres: A comparison of verbal and nonverbal cues in news and thread summaries. In *The 41st Language at Edinburgh Lunch*, Edinburgh, UK.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. Technical report ISITR- 545. Technical report, ISI, University of Southern California.

Cervone, A. (2014). Attribution relations extraction in speech: A lexical-prosodic approach. Master's thesis, Università degli Studi di Pavia, Pavia.

Cervone, A., Pareti, S., Bell, P., Prodanof, I., and Caselli, T. (2014). Detecting attribution relations in speech. In Basili, R., Lenci, A., and Magnini, B., editors, *First Italian Conference on Computational Linguistics CLiC-it 2014*, Pisa, Italy. Pisa University Press.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 132–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 355–362, Morristown, NJ, USA. Association for Computational Linguistics.

Curran, J. R. and Clark, S. (2003). Language independent NER using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, pages 164–167, Morristown, NJ, USA. Association for Computational Linguistics.

Das, D. and Bandyopadhyay, S. (2010). Emotion holder for emotional verbs: The role of subject and syntax. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 385–393. Springer, Berlin, Heidelberg.

de La Clergerie, E., Sagot, B., Stern, R., Denis, P., Recource, G., and Mignot, V. (2009). Extracting and visualizing quotations from news wires. In *Proceedings of L&TC 2009, Poznan, Poland.*

Diab, M., Levin, L., Mitamura, T., Rambow, O., Prabhakaran, V., and Guo, W. (2009). Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 68–73.

Dinesh, N., Lee, A., Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2005). Attribution and the (non-)alignment of syntactic and discourse arguments of connectives. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II:*

Pie in the Sky, CorpusAnno '05, pages 29–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Doandes, M. (2003). Profiling for belief acquisition from reported speech. Master's thesis, Concordia University, Montreal, Quebec, Canada.

Duboue, P. A. (2012). Extractive email thread summarization: Can we do better than he said she said? In *Proceedings of the Seventh International Natural Language Generation Conference*, INLG '12, pages 85–89, Stroudsburg, PA, USA. Association for Computational Linguistics.

Elson, D. K. and McKeown, K. R. (2010). Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10).*

Evans, D. K., Ku, L.-W., Seki, Y., Chen, H.-H., and Kando, N. (2007). Opinion analysis across languages: An overview of and observations from the NTCIR-6 opinion analysis pilot task. In Masulli, F., Mitra, S., and Pasi, G., editors, *Applications of Fuzzy Sets Theory*, volume 4578 of *Lecture Notes in Computer Science*, pages 456–463. Springer, Berlin, Heidelberg.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Fellbaum, C. (1998). *WordNet : An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Fernandes, W., Motta, E., and Milidiú, R. (2011). Quotation extraction for Portuguese. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL 2011)*, pages 204–208, Cuiaba.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Morristown, NJ, USA. Association for Computational Linguistics.

Glass, K. and Bangay, S. (2007). A naive, salience based method for speaker identification in fiction books. In *Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 07).*

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining software: An update. *SIGKDD Exploration Newsletter*, 11(1):10–18.

He, H., Barbosa, D., and Kondrak, G. (2013). Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria.

Hollingsworth, B. and Teufel, S. (2005). Human annotation of lexical chains: Coverage and agreement measures. In *ELECTRA Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications (Beyond Bag of Words)*, pages 26–32.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python.

Joyce, J. (2001). *Ulysses: A Reproduction of the 1922 First Edition.* Dover Publications.

Kessler, J. S. (2008). Polling the blogosphere: A rule-based approach to belief classification. In *International Conference on Weblogs and Social Media (ICWSM).*

Kim, S.-M. and Hovy, E. (2005). Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains, Pennsylvania, US*, Pennsylvania, US,.

Kim, S.-M. and Hovy, E. (2006a). Extracting opinions, opinion holders, and topics expressed in online news media text. In *SST '06: Proceedings of the Workshop on Sentiment and Subjectivity in Text*, Morristown, NJ, USA. Association for Computational Linguistics.

Kim, S.-M. and Hovy, E. (2006b). Identifying and analyzing judgment opinions. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 200–207, Morristown, NJ, USA. Association for Computational Linguistics.

Kim, Y., Jung, Y., and Myaeng, S.-H. (2007). Identifying opinion holders in opinion text from online newspapers. In *Proceedings of the 2007 IEEE International Conference on Granular Computing*, GRC '07, pages 699–702, Washington, DC, USA. IEEE Computer Society.

Kingsbury, P., Palmer, M., and Marcus, M. (2002). Adding predicate argument structure to the Penn TreeBank. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 252–256, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kipper, K., Dang, H. T., and Palmer, M. (2000). Class-based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. AAAI Press.

Klebanov, B. B., Beigman, E., and Diermeier, D. (2010). Vocabulary choice as an indicator of perspective. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 253–257, Stroudsburg, PA, USA. Association for Computational Linguistics.

Klein, D. and Manning, C. D. (2002). Fast exact inference with a factored model for natural language parsing. In *Advances in neural information processing systems*, pages 3–10.

Krestel, R., Bergler, S., and Witte, R. (2008). Minding the source: Automatic tagging of reported speech in newspaper articles. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Krestel, R., Witte, R., and Bergler, S. (2007). Creating a fuzzy believer to model human newspaper readers. In Kobti, Z. and Wu, D., editors, *Advances in Artificial Intelligence*, volume 4509 of *Lecture Notes in Computer Science*, pages 489–501. Springer, Berlin, Heidelberg.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 497–506, New York, NY, USA. ACM.

Levin, B. (1993). *English verb classes and alternations*. University of Chicago Press.

Li, H., Cheng, X., Adson, K., Kirshboim, T., and Xu, F. (2012). Annotating opinions in German political news. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Liang, J., Dhillon, N., and Koperski, K. (2010). A large-scale system for annotating and querying quotations in news feeds. In *Proceedings of the 3rd International Semantic Search Workshop*, SEMSEARCH '10, New York, NY, USA. ACM.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2014). A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(02):151–184.

Lu, B. (2010). Identifying opinion holders and targets with dependency parser in Chinese news texts. In *HLT '10: Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 46–51, Morristown, NJ, USA. Association for Computational Linguistics.

Mamede, N. and Chaleira, P. (2004). Character identification in children stories. *Advances in Natural Language Processing*, pages 82–90.

Mann, W. C. and Thompson, S. A. (1988). *Rhetorical Structure Theory: A theory of text organization*. Ablex.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M. T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F., and Delmonte, R. (2003). Building the Italian syntactic-semantic treebank. In *Abeillé (Abeillé, 2003), chapter 11*, pages 189–210. Kluwer.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun, S., Kohn, K., and Mukherjee, J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Germany.

Murphy, A. C. (2005). Markers of attribution in English and Italian opinion articles: A comparative corpus-based study. *ICAME Journal*, 29:131–150.

Okazaki, N. (2007). CRFsuite: a fast implementation of conditional random fields (CRFs).

O'Keefe, T., Pareti, S., Curran, J., Koprinska, I., and Honnibal, M. (2012). A sequence labelling approach to quote attribution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Jeju, Korea.

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

Pardo, T., das Graças Volpe Nunes, M., and Rino, L. (2004). Dizer: An automatic discourse analyzer for Brazilian Portuguese. In Bazzan, A. and Labidi, S., editors, *Advances in Artificial Intelligence – SBIA 2004*, volume 3171 of *Lecture Notes in Computer Science*, pages 224–234. Springer, Berlin, Heidelberg.

Pardo, T. and Nunes, M. (2003). A construção de um corpus de textos cientficos em português do Brasil e sua marcação retórica. Technical report, São Carlos-SP.

Pareti, S. (2009). Towards a discourse resource for Italian: Developing an annotation schema for attribution. Master's thesis, Università degli Studi di Pavia, Pavia.

Pareti, S. (2011). Annotating attribution relations and their features. In Alonso, O., Kamps, J., and Karlgren, J., editors, *ESAIR'11: Proceedings of the CIKM'11*

Workshop on Exploiting Semantic Annotations in Information Retrieval. ACM Press.

Pareti, S. (2012a). A database of attribution relations. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Pareti, S. (2012b). The independent encoding of attribution relations. In *Proceedings of the Eight Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-8)*, Pisa, Italy.

Pareti, S. (2015). Annotating attribution relations across languages and genres. In *Proceedings of the Eleventh Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK.

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R., and Koprinska, I. (2013). Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA. Association for Computational Linguistics.

Pareti, S. and Prodanof, I. (2010). Annotating attribution relations: Towards an Italian discourse treebank. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of LREC*, Malta. European Language Resources Association (ELRA).

Paroubek, P., Pak, A., and Mostefa, D. (2010). Annotations for opinion mining evaluation in the industrial context of the doxa project. In *Proceedings of the 7th International Conference on Language Resources Evaluation (LREC)*, Malta.

Pouliquen, B., Steinberger, R., and Best, C. (2007). Automatic detection of quotations in multilingual news. In *Proceedings of the International Conference Recent Advances In Natural Language Processing (RANLP 2007)*, pages 487–492.

Prasad, R., Dinesh, N., Lee, A., Joshi, A., and Webber, B. (2006). Annotating attribution in the Penn Discourse TreeBank. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 31–38.

Prasad, R., Dinesh, N., Lee, A., Joshi, A., and Webber, B. (2007). Attribution and its annotation in the Penn Discourse TreeBank. *TAL (Traitement Automatique des Langues)*, 42(2).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation LREC*, pages 2961–2968.

Pustejovsky, J., Castao, J., Ingria, R., Saur, R., Gaizauskas, R., Setzer, A., and Katz, G. (2003a). TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of IWCS-5, Fifth International Workshop on Computational Semantics*.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003b). The timebank corpus. In *Corpus linguistics*, pages 647–656.

Pustejovsky, J., Littman, J., Saurí, R., and Verhagen, M. (2006). Timebank 1.2 documentation. Technical report.

Ramshaw, L. A. and Marcus, M. P. (1995). Text chunking using transformation-based learning. In *Proceedings of the Third Annual Workshop on Very Large Corpora (ACL)*, pages 82–94.

Renzi, L., Salvi, G., and Cardinaletti, A. (1995). *Grande Grammatica Italiana di Consultazione*, volume III, pages 431–436. Il Mulino, Bologna.

Ruppenhofer, J., Sporleder, C., and Shirokov, F. (2010). Speaker attribution in cabinet protocols. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC*, Malta. European Language Resources Association (ELRA).

Sag, I. A. and Pollard, C. (1991). An integrated theory of complement control. *Language*, 67(1):63–113.

Sams, J. (2008). *A Syntactic Analysis of Written English Quotatives*. PhD thesis, University of Colorado at Boulder.

Santos, C. and Milidiú, R. (2009). Entropy guided transformation learning. In Hassanien, A.-E., Abraham, A., Vasilakos, A., and Pedrycz, W., editors, *Foundations of Computational Intelligence Volume*, volume 1, pages 159–184. Springer, Berlin, Heidelberg.

Sarmento, L. and Nunes, S. (2009). Automatic extraction of quotes and topics from news feeds. In *Proceedings of DSIE'09 - 4th Doctoral Symposium on Informatics Engineering*.

Saurí, R. and Pustejovsky, J. (2009). Factbank: A corpus annotated with event factuality. *In Language Resources and Evaluation*, (43):227–268.

Schneider, N., Hwa, R., Gianfortoni, P., Das, D., Heilman, M., Black, A. W., Crabbe, F. L., and Smith, N. A. (2010). Visualizing topical quotations over time to understand news discourse. Technical Report T.R. CMU-LTI-10-013, Carnegie Mellon University, Pittsburgh, PA.

Schuler, K. K. (2005). *Verbnet: a broad-coverage, comprehensive verb lexicon*. PhD thesis, Philadelphia, PA, USA.

Seki, Y., Evans, D. K., Ku, L.-W., Sun, L., Chen, H.-H., and Kando, N. (2008). Overview of multilingual opinion analysis task at NTCIR-7. In *Proceedings of NTCIR-7 Workshop Meeting on Evaluation of Information Access Technologies*, pages 185–203, NII, Japan.

Seki, Y., Ku, L.-W., and Sun, L. (2010). Overview of multilingual opinion analysis task at NTCIR- 8. In *Proceedings of NTCIR-8 Workshop Meeting on Evaluation of Information Access Technologies*, pages 209–220, Tokyo, Japan.

Skadhauge, P. R. and Hardt, D. (2005). Syntactic identification of attribution in the RST treebank. In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora*, Jeju Island, Korea.

Thompson, G. and Yiyun, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, 12(4):365–382.

van Leeuwen, T. (2008). *Discourse and Practice: New Tools for Critical Analysis*, chapter 2 Representing Social Actors, pages 24–54. Oxford University Press, New York.

Watterson, B. (1994). *Homicidal Psycho Jungle Cat: A Calvin and Hobbes Collection*. Andrews McMeel Publishing, 1 edition.

Weischedel, R. and Brunstein, A. (2005). BBN pronoun coreference and entity type corpus. *Linguistic Data Consortium*.

Weiser, S. and Watrin, P. (2012). Extraction of unmarked quotations in newspapers. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Wiebe, J. (2002). Instructions for annotating opinions in newspaper articles. Technical report, University of Pittsburgh, Pittsburgh, PA.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

Wiegand, M. and Klakow, D. (2010). Convolution kernels for opinion holder extraction. In *HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 795–803, Morristown, NJ, USA. Association for Computational Linguistics.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31:249–288.

Xu, R., Wong, K.-F., Lu, Q., Xia, Y., and Li, W. (2008). Learning knowledge from relevant webpage for opinion analysis. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT '08*, volume 1, pages 307–313.

Zhang, J., Black, A., and Sproat, R. (2003). Identifying speakers in children's stories for speech synthesis. In *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland. ISCA.