Annotating Attribution Relations: Towards an Italian Discourse Treebank

Silvia Pareti*, Irina Prodanof+
*Università degli Studi di Pavia, C.so Strada Nuova 65, 27100 Pavia
+Istituto di Linguistica Computazionale (ILC), CNR, Via Moruzzi 1, 56124 Pisa
E-mail: silvia_prt@yahoo.it, irina.prodanof@ilc.cnr.it

Abstract

In this paper we describe the development of a schema for the annotation of attribution relations and present the first findings and some relevant issues concerning this phenomenon. Following the D-LTAG approach to discourse, we have developed a lexically anchored description of attribution, considering this relation, contrary to the approach in the PDTB, independently from other discourse relations. This approach has allowed us to deal with the phenomenon in a broader perspective than previous studies, reaching therefore a more accurate description of it and making it possible to raise some still unaddressed issues. Following this analysis, we propose an annotation schema and discuss the first results concerning its applicability. The schema has been applied to a pilot portion of the ISST corpus of Italian and represents the initial phase of a project aiming at the creation of an Italian Discourse Treebank. We believe this work will raise some awareness concerning the fundamental importance of attribution relations. The identification of the source has in fact strong implications for the attributed material. Moreover, it will make overt the complexity of a phenomenon for long underestimated.

1. Introduction

In this paper we present an annotation schema for attribution relations and discuss the major issues concerning the phenomenon. An increasing number of discourse corpora have been recently developed, allowing significant progress in this field. However, only few of these resources have included attribution, although only partially, in their account of discourse phenomena. Inspired by the work in the PDTB (Prasad et al., 2007) and the Opinion Corpus (Wilson and Wiebe, 2005), we have developed an annotation schema for attribution independent from other discourse phenomena and annotated attribution relations on a portion of the ISST[1] in order to assess its feasibility. We believe this kind of resource represents a step towards covering a gap in the field of discourse analysis and would provide useful material to develop and test advanced systems able to discern information on the basis of its provenance. Different sources differ in bias and reliability and this can affect the way information is perceived and should be handled. This is particularly relevant for studies in the fields of Information Retrieval, Information Extraction, MPQA, and Opinion Mining.

We will proceed first with defining the scope of our study, the framework from which it originates and the components of the attribution relation (2). In (3) we present the attributes and relative values included in the proposed annotation schema. Afterwards (4), we describe the pilot annotation project and (5) the distribution of the different attribute values and some possible interactions. Finally, we discuss some relevant issues concerning attribution (6) and conclude (7) with the future work we intend to undertake.

2. Attribution

The concept of attribution in a text has been considered in this study as the relation ascribing the ownership of an attitude towards some linguistic material, i.e. the text itself, a portion of it or their semantic content, to an entity. This definition allows considering instances of attribution going beyond the sentence boundaries, unlike the Opinion Corpus (Wilson and Wiebe, 2005), and attributed material consisting e.g. of a single word, and not just Abstract Objects[2] as in the PDTB (Prasad et al., 2007). The PDTB has been nonetheless our starting point for the annotation of attribution relation. This because, while works in the frame of Opinion Mining focus especially on the kind and polarity of the opinion itself, the annotation schema proposed in the PDTB appeared as more suitable in order to address a wider spectrum of attribution relations. Their annotation comprises relevant attributes, peculiar to attribution, e.g. the type of source. However, in the PDTB attribution relations are annotated only when overlapping with a relation conveyed by a discourse connective. The attributed material can therefore be the discourse connective itself, or one of its arguments (Arg1, Arg2). Since this approach leaves out several instances of attribution and therefore some related issues, e.g. nested attribution, their schema had to be adapted to suit our task.

In order to accomplish reaching a broader account of attribution, thus enabling a better handling of the phenomenon, not only the annotation is independent from other discourse relations, but any kind of attributed material has been considered. This can be a single word as well as one or more phrases, clauses, sentences or the entire document, i.e. the article itself.

In the account of attribution we have adopted a lexicalized perspective similar to the D-LTAG approach to discourse (Cristea and Webber, 1997, Webber et al., 2003). Therefore this particular discourse relation has also been lexically anchored, resulting in attribution been analysed as composed by three elements. According to the terminology adopted, these are: the content, i.e. the

[1] ISST- Italian Syntactic-Semantic Treebank (Montemagni et al., 2003).

[2] Abstract Objects are propositions, events or states.

attributed material; the source, that is the entity the material is attributed to; and the cue. This last element represents the lexical material signalling the attribution and functions as a link between source and content. These markables do not completely match the spans annotated in the Opinion Corpus (Wiebe, 2002), i.e. inside, outside and on. The source is in fact annotated independently in our schema, while in the Opinion Corpus this is labelled as 'outside' together with everything in the sentence other than the 'on' (cue) and 'inside' (content) as in the example below (Wiebe, 2002:8):

(a) **outside**: "On Tuesday, John …while hanging up the phone."
on: "said that"
inside: "he was leaving"

In order to account for material which could not be incorporated in any of the three constitutive elements of the attribution relation but nonetheless was significant for the interpretation of the content, an optional fourth markables, the supplement, has also been added to the schema.
Several elements have been identified which can assume one of the above mentioned roles, as exemplified in Table 1 below. The analysis is tailored on Italian, however it generally holds also for other languages with some modifications. The English language, for example, cannot express the cue making use of a grammatical marker. The 'quotative conditional'[3] is in fact prerogative of Italian, and some other languages such as French. On the other hand, in English, unlike Italian, adverbs (e.g. allegedly, reportedly) can also function as attribution anchor.

Source	*noun phrase, adjective, prep. phrase*
Cue	*verb, noun, adjective, preposition, prep. group, (grammatical marker), graphic marker*
Content	*word, phrase, clause, sentence, article*
Supplement	*cue modifier, indirect object, source of source, event specification*

Table 1 - Markables and elements expressing them

3. Annotation Schema

The different constituents of attribution, i.e. source, cue, content and supplement, represent the core of the annotation and are labelled as individual markables. A collection of guidelines has been prepared in order to provide advice on what to include in each markable span. In order to make the interconnection of these elements explicit, the markables being part of the same attribution have been linked grouping them in a <relation>. Each relation has one cue markable, at least one content markable and none or more source and supplement markables. Unlike the supplement, the source is not optional, however, in a pro-drop language like Italian, this might not appear in the text and be left implicit. In addition to the markables, relevant features have also been annotated and arbitrarily marked on the cue.
These features are a modification of the features included in the annotation of attribution in the PDTB. They provide preliminary information about the source, the type of attitude, the factuality of the attribution and the eventual change in the scope of an element affecting the factuality. The values they can assume are exemplified in Table 2.
The <source> attribute can assume the values: WRITER, left implicit unless explicitly stated in the text; OTHER, to refer to a specific entity; ARBITRARY, when the attribution refers to a general or not specified entity (e.g. hearsay); and MIXED, in case of a multiple source of dissimilar kind.

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 - Annotation schema attributes and their values

The <type> can assume four values, namely ASSERTION (expressing a communicative act), BELIEF (reflecting a mental attitude), FACT (conveying knowledge perception or possession) and EVENTUALITY (dealing with intentions). The <factuality> expresses whether the attribution relation is only supposed, possible or unreal or it is presented as a fact of the real world (e.g. *John could think* vs. *John thinks*). It is not making any judgement on the factuality of the content itself. Elements affecting the factuality are:

- polarity reversing particles (e.g. negation, negative pronouns)
- verb mode (e.g. conditional, imperative)
- verb tense (e.g. future)
- hypothetical (e.g. if)
- interrogative form
- modals

These elements usually affect directly the cue, but occasionally they can be found on the source, as in the example below, where the indefinite pronoun results in an empty source:

(b) **Nessuno** parla più *di baratro imminente e di crisi finanziaria.* (ISST cs025)
No one is talking anymore *about imminent precipice and financial crisis.*

With <scopal_change> are marked those instances of attribution appearing as NON-FACTUAL on the surface while it is actually the content and not the cue that is

[3] e.g. 'Un incendio, che si sarebbe sviluppato:COND per cause accidentali, ...'(ISST cs010) / *A fire, which (is said to have) developed for accidental causes, ...*

affected e.g. 'Non credo sia una buona idea'/ *I don't think it is a good idea = I think it is not a good idea*.
With respect to the PDTB scheme, the one proposed above introduces some changes, partly due to constraints determined by the tool characteristics, partly because of new features of the phenomenon emerging from the preliminary analysis. Since it was not possible to assign different values to the attribute <source>, we added the value MIXED, in order to account for multiple sources comprising different types, e.g. ARBITRARY and OTHER as in (c). Multiple sources of the same kind, e.g. *John, Mary and Peter*, have been instead grouped into the same markable.

(c) **Tutti**, incluse **le autorità**, conoscono *la loro provenienza*, ma nessuno dice e fa nulla per prevenire il massacro di capi selvatici. (ISST cs020)
Everyone, including **the authorities**, knows *their provenance*, but no one says and does anything to prevent the massacre of wild animals.

On the conceptual level, we have renamed as <factuality> the attribute 'determinacy', since this seems a more appropriate label, and changed 'scopal polarity' in <scopal_change>. From the analysis emerged in fact that elements whose change in scope could affect the factuality of the attribution are not just polarity modifiers. The conjunction 'if', for example, can also superficially scope on the cue and affect instead the content. In (d), although the first part of the sentence apparently is a condition for 'think' to happen, it is already part of the content and scopes inside it. The whole sentence could be rewritten as: *I think that, if there is a majority…, the legislature could usefully continue*.

(d) *Se c'è, cioè, una maggioranza in Parlamento in grado di affrontare seriamente una fase di riforme anche elettorali,* **Ø** penso che *la legislatura possa utilmente proseguire*.' (ISST re075)
If there is a majority at the Parliament able to seriously face a phase of reforms, also electoral, **(I)** think that *the legislature could usefully continue*.

Further investigations are required to identify which elements can present a change in scope and derive possible constraints to the combination of features in the schema.

4. Building the Corpus

The corpus that has been chosen for the addition of this level of annotation is the ISST corpus (Montemagni et al., 2003), which consists of 307.682 word tokens from a collection of 484 articles drawn from Italian newspapers and periodicals. The reason beneath this choice is that this corpus already represents a complete resource, since it encodes, in separate levels of annotation, orthographic, morpho-syntactic, syntactic and semantic information. Moreover, the ISST is comparable to the PDTB corpus, from which the present annotation schema was derived, since both consist of news articles and are representative of the newspaper language.
The pilot corpus annotated for attribution comprises 50 articles drawn from the ISST corpus, selected in order to obtain a balanced subcorpus. The overall number of tokens is 37.000. Considering the pervasiveness of the phenomenon in journalistic language, the size of the corpus can be considered already significant. In the pilot corpus in fact an overall number of 461 attribution relations have been identified and annotated, that is an average of 9,22 relations per article. Although the study will benefit from the extension of the annotation to the whole ISST corpus, all the different attributes included in the annotation schema are already represented in the pilot. This has allowed a complete verification of the schema applicability.
The tool adopted and tailored for the annotation was MMAX2[4] (Mueller and Strube, 2006). This was chosen after an in depth comparison of several available tools (e.g. GATE, Knowtator, Callisto) as it best supports the specific annotation requirements determined by the peculiarities of the phenomenon. These include: the annotation of discontinuous and multiple text spans as a single markable; the possibility of establishing relations among two or more markables; the annotation of overlapping markables. At this stage, the annotation was performed manually by a single annotator.

5. Attribution Figures

We report in this chapter some figures and tables concerning the distribution of attribution values in the pilot corpus. Although drawn from a relatively small corpus, some tendencies and attribute correlations are already visible in the data. This preliminary survey of attribution does not intend to provide definitive results but to present an initial overview of this discourse relation. The observation of the data gives an idea of the proportion of the phenomenon and can already suggest possible issues.
In the corpus, an overall number of 461 relations were detected and annotated. Since in our annotation every relation has one and only one cue, this also corresponds to the overall number of cue markables. On the other hand, source markables are just 329 meaning that 132 sources, i.e. more than a fourth, are not explicitly mentioned. This would be an extremely high number with relevant implications, considering for example that studies in the area of automatic detection of opinion sources have generally not addressed implicit sources recognition (Choi et al., 2006). This because they are based on the Opinion Corpus where the occurrence of implicit sources is relatively low, about 7%, and therefore a minor issue. Although it is possible that this percentage is doomed to rise when considering all type of attributions, also beyond the sentence boundaries, the high percentage of missing sources in our corpus is mainly due to the peculiarity of the Italian language, i.e. subject pro-drop. Most of the attribution relations that do not have a corresponding source markable are of this nature. About half of the overall number of implicit sources are of the type ARBITRARY. Moreover, they also account for about half the sources of NON-FACTUAL attributions.

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

Content markables on the other hand, are slightly more than the overall number of relations. They sum up to just 468 markables, however, multiple contents are a very common feature. This small number is due to the fact that it has been chosen to annotate multiple contents as a single markable also when not adjacent, unless they are also separate by sentence boundaries.
Table 3 shows the distribution of the different types of relation, namely the attitude the source holds towards the content. Considering the corpus is composed of newspaper articles, it is not surprising that the attribution of communicative acts, i.e. ASSERTION, is by far the most frequent. What is interesting to notice is instead the correlation of the attribute <type> with the <factuality> attribute. NON-FACTUAL occurs on average in about a tenth of the attributions, however, it represents almost a third of the relations of the FACT and EVENTUALITY type. This suggests that unreal or just supposed attributions tend to occur a lot more frequently when talking about other people's knowledge or intentions than when reporting their assertions or beliefs.

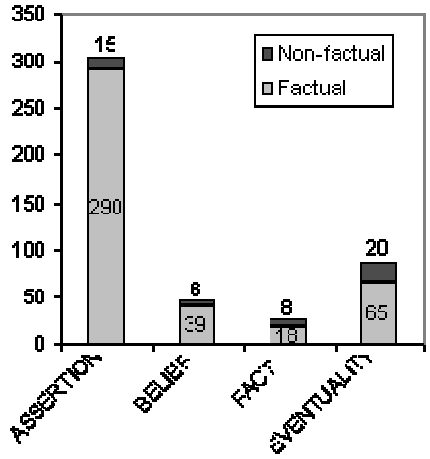

Table 3 - Value distribution of the attribute <type>

An attribution of the FACT type can also be NON-FACTUAL as in the example (e) since the <type> reflects the attitude the source holds, while the <factuality> the truth value of the attribution relation itself.

(e) **Nessuno** sa ancora, *se la Berlino del Duemila davvero sarà una Parigi tedesca, o solo una riedizione abbellita e troppo maestosa della città frammentata, vivace e violenta di oggi*. (ISST re001)
No one knows yet, *if the Berlino of the 21st century is really going to be a German Paris, or just an embellished and too majestic republication of the fragmented, lively and violent city of today*.

Concerning the source (Table 4), the relation has in 375 cases the value OTHER, which means the content is attributed to a specific entity. Although this is by far the predominant type of source, attributions to an ARBITRARY source or explicitly to the WRITER are also relatively common and need to be accounted for. The value MIXED occurred instead only a single time since there was only one instance of multiple source comprising different types of attribution holders (example (c) above).

WRITER	23
OTHER	375
ARBITRARY	62
MIXED	1

Table 4 - Value distribution of the attribute <source>

The feature <scopal_change> is a very marginal phenomenon in the corpus. There are in fact only 7 instances of a change in scope (Table 5). The attribution relations of the type FACT have not been annotated for this feature, since they are not affected by it (Kiparsky and Kiparsky, 1971). All the instances of SCOPAL CHANGE co-occur in the corpus with attributions of the EVENTUALITY type.

NONE	429
SCOPAL-CHANGE	7

Table 5 - Value distribution of the attribute <scopal_change>

6. Final Remarks

Several issues emerged during the analysis and while performing the pilot annotation which contribute to making attribution a very complex phenomenon. The most interesting matters are here briefly introduced. One aspect is the necessity to preliminary perform coreference resolution as the two phenomena are strongly intertwined and the referents of anaphoric components of attribution need to be first retrieved for the relation to be informative. In (f) for example the content of the first attribution is not expressed but recalled with the pronoun 'it', while the source of the second relation is the relative pronoun 'who'.

(f) *Lo* ha detto ieri **un portavoce del ministero degli Esteri**, **il quale** ha anche annunciato che *il governo cinese ha protestato con quello degli Stati Uniti e che si riserva il diritto di ulteriori reazioni*. (ISST els075)
It was said yesterday **by a spokesman of the Foreign Ministry**, **who** has also announced that *the Chinese government has complained to the one of the United States and that they reserve themselves the right of further reactions.*

We also found a correlation between the kind of anaphora and the element in the relation it substitutes. Sources are often recalled by pronominal or bridging anaphora, while the content relies mainly on event anaphora (when attributed are e.g. *discourse, press release, words*).
Another pervasive aspect, that has not been considered in the PDTB, is the nesting of attribution relations one in another. The content of an attribution can have inside another attribution relation, often reaching several levels of embedding. The practice of 'recycling' information is quite widespread as newspapers often report news acquired from other intermediary. It is crucial to determine all the passages the information has undergone, i.e. all the sources of an embedded content, in order to

make judgements concerning their trustworthiness and to interpret the content itself.

(g) **Blinder**, secondo **voci** riferite dal **New York Times**, sperava di succedere al presidente Greenspan quando a marzo scadrà la sua nomina. (ISST re070)
Blinder, according to **rumours** reported by **the New York Times**, hoped to succeed to president Greenspan when in May his appointment will run over.

While the Opinion Corpus lists all the sources in the 'source_ID' slot of each attribution, we consider this task avoidable, as the embedding could be automatically derived considering the inclusion of an annotated span into another.
Concerning the attitude <type>, the classification adopted from the PDTB presents some difficulties, as brought up by the annotation. Cues composed of contrasting elements, together with multimodal verbs, e.g. 'Arlacchi sorride: "…"' (ISST re095) / *Arlacchi smiles: "…"*, contribute to making the task of assigning the type particularly complex for the annotator. We prepared detailed information on how to perform this task, which needs to be evaluated confronting interannotator agreement score, possibly leading to some modifications to the scheme. We intend to extract from the annotated corpus a list of possible cues. These could also be grouped according to the <type>, however, it is not possible to compile a predefined set, since most verbs are polysemous and their type can only be determined in context as in the following examples where the same verb is an ASSERTION in (h) and an EVENTUALITY in (i):

(h) 'Il governo di Zagabria, invece, sostiene che sono "solo" 100 mila le persone in cammino.' (ISST cs031)
Zagreb government claims instead that they are only 100 thousand the people who set out.

(i) 'Ma ieri sera I parlamentari serbi hanno "sostenuto senza riserve" la decisione di Karadzic'. (ISST cs034)
However yesterday evening the Serbian parliamentarians have "supported wholeheartedly" Karadzic's decision.

A last issue presented here is the presence of elements, identified with 'source of source', which can represent either an additional source of FACT attributions as in (j), or an intermediary source of an ASSERTION as in the example (k), implicitly suggesting the existence of an additional level of embedding of the content, i.e. a spokesman. These elements have been annotated as 'supplement', hence making it possible to retrieve them at a later stage.

(j) (Ø) Ho saputo della squalifica di Garciano da Maurizio Damilano, vi giuro, non pensavo di arrivare primo. (ISST cs071)
(I) heard of the disqualification of Garciano from Maurizio Damilano, I swear, I didn't imagine I would have came first.

(k) Poi però, tramite la figlia che sta a Santiago, prima (Ø) limita la portata del colloquio con Gaston Salvatore ("non è stata una vera intervista, solo una conversazione"), poi (Ø) smentisce. (ISST period005)
Afterwards however, through the daughter who lives in Santiago, first (she) diminishes the importance of the colloquium with Gaston Salvatore ("it wasn't a real interview, just a conversation"), then (she) denies.

7. Conclusion and future work

The study presented in this paper is still to be completed. Some aspect of attribution require further investigation and the annotation schema needs to be tested for interannotator agreement[5] before being applied to the whole ISST corpus. However, we believe that the results achieved are already significant and could inspire similar projects. The study has in fact achieved:

- a deep and independent analysis of attribution relations;
- the definition of a complete annotation schema for attribution;
- the construction of a small corpus annotated for attribution relations which represents a resource available for other studies;
- the identification of issues and aspects concerning attribution and its annotation.

Although the analysis of attribution will still hold, genres other than newspaper language, such as juridical or political reports, literature or dialogue, surely present a different distribution of the phenomenon and peculiarities which need to be considered when dealing with the identification of the source. In particular, it would be interesting to study the phenomenon in dialogue language, for example by applying the schema to the LUNA[6] corpus. This would be particularly significant since a preliminary study of the applicability of the PDTB schema and the annotation of relations conveyed by discourse connectives on a portion of this corpus has already been performed (Tonelli et al., 2010). Being this a multilingual corpus, this would also enable inter-linguistic comparison of attribution in dialogues.

8. Acknowledgements

The project in this paper was mainly developed at the School of Informatics of the University of Edinburgh, thanks to an Erasmus Placement grant. Among the many fruitful ones, we would like to particularly acknowledge the contributions of Bonnie Webber and Rashmi Prasad who provided both technical support and continuous suggestions along all the development stages.

9. References

Choi, Y., Breck, E., Cardie, C. (2006). Joint extraction of entities and relations for opinion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[5] Although this is a priority, it has not yet been possible to find and train the workforce to perform this task.
[6] 1EU FP6 contract No. 33549, http://www.ist-luna.eu/

Cristea, D., Webber, B. (1997). Expectations in Incremental Discourse Processing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, pp. 88--95.

Kiparsky, C, Kiparsky, P. (1971). Fact. In Jakobovits, L., Steinberg, D. (Eds.), *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*. Cambridge: Cambridge University Press, pp.345--369.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M. T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F., Delmonte, R. (2003). Building the Italian Syntactic-Semantic Treebank. In Abeillé, A. (Eds.), *Building and using Parsed Corpora, Language and Speech series*. Kluwer, Dordrecht, pp. 189--210.

Mueller, C., Strube, M. (2006). Multi-level Annotation of Linguistic Data with MMAX2. In Braun, S., Kohn, K., Mukherjee, J. (Eds.), *Corpus Technology and Language Pedagogy*. *New Resources, New Tools, New Methods*, (English Corpus Linguistics, vol.3), Frankfurt: Peter Lang, pp.197--214.

Pareti, S. (2009). Towards a Discourse Resource for Italian: Developing an Annotation Schema for Attribution. Thesis of the Master in Theoretical and Applied Linguistics. Università degli Studi di Pavia, Pavia.

Prasad, R., Dinesh, N., Lee, A., Joshi, A., Webber, B. (2007). Attribution and its Annotation in the Penn Discourse TreeBank. In *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse*, 47(2), pp. 43--64.

Tonelli, S., Riccardi, G., Prasad, R., Joshi, A. (2010). Annotation of Discourse Relations for Conversational Spoken Dialogues. To appear in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. Malta, 17-23 May.

Webber, B., Stone, M., Joshi, A., Knott, A. (2003). Anaphora and Discourse Structure. *Computational Linguistics*, 29, pp. 545--587.

Wiebe, J. (2002). Instructions for annotating opinions in newspaper articles. Technical report TR-02-101, Department of Computer Science, University of Pittsburgh.

Wilson, T., Wiebe, J. (2005). Annotating Attributions and Private States. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. Ann Arbor, Michigan.