Epistemic Modality and Knowledge Attribution in Scientific Discourse: A Taxonomy of Types and Overview of Features

Anita de Waard
Elsevier Labs
Jericho, VT, USA
a.dewaard@elsevier.com

Henk Pander Maat
Utrecht Institute of Linguistics
Utrecht, The Netherlands
h.l.w.pandermaat@uu.nl

Abstract

We propose a model for knowledge attribution and epistemic evaluation in scientific discourse, consisting of three dimensions with different values: source (author, other, unknown); value (unknown, possible, probable, presumed true) and basis (reasoning, data, other). Based on a literature review, we investigate four linguistic features that mark different types epistemic evaluation (modal auxiliary verbs, adverbs/adjectives, reporting verbs and references). A corpus study on two biology papers indicates the usefulness of this model, and suggest some typical trends. In particular, we find that matrix clauses with a reporting verb of the form '*These results suggest'*, are the predominant feature indicating knowledge attribution in scientific text.

1 Introduction

Our main research goal is to linguistically "specify the precise time and place in the process of fact construction when a statement became transformed into a fact", as Latour and Woolgar (1979) put it. Specifically, we are interested in creating a linguistically motivated framework of biological sensemaking to help extract newly claimed knowledge from large text corpora.

Biological understanding consists of a conceptual model of the system at study, which is collaboratively created by the scientists working on that system. In contributing a new building block to the model, authors will need to argue, first: that their experiments are appropriate, and performed well; second, that they can draw certain conclusions from these experiments; and third, that, and how, these conclusions fit within the existing knowledge model for their field. Their observations and inferences might confirm or contradict other thoughts about the model, expressed in other papers. This need to indicate certainty and agreement/disagreement means that biological papers contain many explicit truth evaluations of their own and other authors' propositions (epistemic modality), and where needed, the explicit attribution of the creator of the propositions (knowledge attribution[1]). Therefore, to understand how biological knowledge is formulated in language, it is essential to understand the linguistic mechanisms of modality and attribution.

In this paper, we present an overview of work in linguistics, genre studies, bioinformatics and computational linguistics, related to epistemic evaluation. From this, we distill a three-tiered taxonomy and a set of linguistic cues or markers that distinguish various forms of epistemic evaluation. We try out this taxonomy and marker set in a small manual corpus exploration of two biology papers, and discuss some correlations between different types and market. We conclude with a proposal for the application of this work.

2 Epistemic Evaluation Taxonomy

2.1 Overview of current work

Strictly speaking, every factual proposition or piece of Propositional Content (Hengeveld and Mackenzie, 2008) contains an (implicit) epistemic evaluation: if a statement is given without further comment on its truth value, we read – irony aside – that the author agrees with the proposition it contains. '*Water is wet.'* – or '*LPS-induced IL-6*

[1] To avoid the use of the cumbersome contraction 'epistemic modality evaluation and knowledge attribution' we will henceforth use the term 'epistemic evaluation' to cover both evaluation and attribution.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 47–55,
Jeju, Republic of Korea, 12 July 2012.

gene transcription in murine monocytes is controlled by NF-B' are statements that do not contain any epistemic modifiers, and are therefore read to be unconditionally accepted by the author. In other cases, however, this truth value is modified: '*These results suggest that water is wet.*' or attributed: '*Author X et al. (2010) report that water is wet.*' Here, we investigate modifiers of propositional content that define either epistemic modality, i.e. the degree of authorial commitment to a proposition, e.g. '*5' untranslated exon 1 may have a regulatory function*', or knowledge attribution: the source of the propositional knowledge, such as when a reference indicates the source of the claim: '*GATA-1 transactivates the EOS47 promoter through a site in the 5'UTR [34].*' There is a body of work pertaining to knowledge attribution and epistemic evaluation in scientific text, within at least four different fields: linguistics, genre studies, bioinformatics, and sentiment detection. A detailed overview of the hedging types and markers found in this literature overview is posted in Dataverse (de Waard, 2012) but we will provide a summary here.

Within linguistics, truth evaluations and source attributions are an important subject within most modern theories of language; here, only a small overview of some pertinent theories can be given. Hengeveld and Mackenzie (2008) characterize truth evaluations as 'modifiers of Propositional Content', concerning 'the kind and degree of commitment of a rational being to Propositional Content, or a specification of the (non-verbal) source of the Propositional Content'. These two categories – knowledge evaluation and knowledge attribution- are also indicated by the concepts 'epistemic modality' and 'evidentiality', respectively. De Haan (1999) strongly argues that they are separate phenomena – and we agree – but for our purposes, establishing modes of truth evaluation and attribution in scientific text, both are relevant. Verstraete (2001) distinguishes between objective and subjective modality: in an objectively modal clause, the truth value of the state of knowledge is brought into question ('*This subject is unknown'*), but the certainty the author has pertaining to the clause is not; in a subjective modal clause, the author expresses uncertainty regarding the extent of his or her knowledge ('*It might be (that this is the case)*').

In genre studies, a body of work revolves around the concept of *hedging*: 'the expression of tentativeness and possibility in language' (Lakoff, 1972; Hyland, 1995). The focus here is on the rhetorical/sociological motivation for, and surface features of, these 'politeness markers'. Myers (1992) identifies stereotypical sentence patterns for hedging from a corpus study of fifty related articles in molecular genetics. Salager-Meyer (1994) defines hedging as presenting 'the true state of the writers' understanding, namely, the strongest claim a careful researcher can make.' She identifies three reasons for hedging: (1) that of purposive fuzziness and vagueness (threat- minimizing strategy); (2) that which reflects the authors' modesty for their achievements and avoidance of personal involvement; and (3) that related to the impossibility or unwillingness of reaching absolute accuracy and of quantifying all the phenomena under observation. Very influentially, Hyland (1995, 2005) proposes an explanatory framework for scientific hedging which combines sociological, linguistic, and discourse analytic perspectives and proposes a three-part taxonomy, distinguishing writer-oriented, accuracy-oriented and reader-oriented hedges. Countering Hyland, Crompton (1997) reviews and evaluates some of the different ways in which the term 'hedge' has been defined in the literature thus far. His new definition is that 'a hedge is an item of language, which a speaker uses to explicitly qualify his/her lack of commitment to the truth of a proposition he/she utters.' Martín-Martín (2008) analyses three different hedging strategies and multiple surface features for hedging in a corpus of full-text papers in English and Spanish, and presents a detailed taxonomy of hedging types and cues, based on literature and corpus studies.

Within bioinformatics and bio-computational linguistics, a body of work has been done on identifying 'speculative language' (Light, 2004). The main purpose here is to enable the automated identification of truth and speculation, in order to enable the construction of databases of known, and candidate, biological facts. The differences with earlier discussions are twofold: first, there is less (or no) effort to study communicative functions: for instance, there is no interest in identifying the authors' rhetorical intent, or the sociological or political motivations for using a particular type of hedge. Second, bioinformatics focuses more on

identifying different types of speculation: is the opinion presented positive or negative, strong or weak, etc. Light et al. (2004) annotate a corpus of Medline sentences as highly speculative, low speculative, or definite, and then train a classifier to automatically recognize speculative sentences. (As an interesting result, they find that almost all speculations appear in the final or penultimate sentence of the abstract).

Wilbur et al. (2006) are motivated by the need to identify and characterize locations in published papers where reliable scientific facts can be found, and present a set of guidelines and the results of an annotation task to annotate a full-text corpus with a five-dimensional set of quantities focus, polarity, certainty, evidence, and directionality. Of these, certainty and evidence relate to knowledge attribution and epistemic evaluation. Medlock and Briscoe (2007) develop a set of guidelines for identifying speculative sentences and an annotated corpus, to test their automated speculation classification tool. Kilicoglu and Bergler (2008) explore a linguistically motivated approach to the problem of recognizing speculative language in biomedical research articles. Building on Hyland's work, they identify **a set of syntactic patterns, which they use for detecting speculative sentences out of a corpus. Thompson et al. (2008) propose a multi-dimensional classification of** a preliminary set of words and phrases that express modality within biomedical texts, and present the results of an annotation experiment where sentences are annotated with level of speculation, type/source of the evidence and the level of certainty towards the statement writer or other. Vincze et al. (2008) describe the BioScope corpus, a collection of Medline abstracts and four full-text papers annotated with instances of negation and speculation.

In the subfield of computational linguistics pertaining to sentiment detection, the goal has been to create overviews of large set of documents summarizing collective opinions and emotion about some topic. Here a more 'mathematical' definition of modality is evolving, which considers the proposition being evaluated as being 'operated on' by the evaluator. A distinction is made between the holder of the opinion, and the strength, polarity and other attributes of the opinion. Similar to work in (bio)computational linguistics, this work has focused is on different types of opinions , and the clues that allow automated detection. Most work in this field has focused on other domains, such as news and product reviews, see e.g. Wilson and Wiebe (2003), Kim and Hovy (2004), and Tang et al., (2009).

2.2 Our proposal

Following the formalism used in opinion/sentiment analysis (e.g., Wilson and Wiebe, 2003; Hovy, 2011) and Functional Discourse Grammar (Hengeveld and Mackenzie, 2008) we differentiate between, firstly, Propositions (similar to FDG's Propositional Content), which can consist of either experimental ('*all thymocytes stained positive for GFP*') or conceptual ('*CCR3 is expressed strongly on eosinophils'*) statements about the (conceived or acted upon/perceived) world, and secondly, modifiers, that modify on these Propositions and modify their truth value or the knowledge attribution. Building on the literature as summarized above, we define a taxonomy of epistemic evaluation along three facets:

1. Epistemic valuations possess a value or level of certainty. Both Hengeveld and Mackenzie (2008) and Wilbur et al. (2006) propose a tripartite division:
   - 'Doxastic' (firm belief in truth, Wilbur's category 3)
   - 'Dubitative' (some doubt about the truth exists; Wilbur's category 2)
   - 'Hypothetical' (where the truth value is only proposed; Wilbur's category 1)
   - Wilbur also adds the useful category 'Lack of knowledge' (level 0).

2. There can different bases of the evaluation:
   - Reasoning: based mostly or solely on argumentation, and not directly on data (e.g., '*it is thought that*', '*we expected*')
   - Data: based explicitly on data (e.g., '*these data suggest that'*, '*CCR3 has been shown to be'*)
   - Implicit or absent: if it is unclear what the evaluation or attribution is based on (e.g., '*GATA-1 transactivates the EOS47 promoter, through a site in the 5'UTR')*

3. The source of the knowledge is identified:
   - Explicit source of knowledge: the knowledge evaluation can be explicitly

owned by the author ('*We therefore conclude that…*') or by a named referent ('*Vijh et al. [28] demonstrated that…*')

- Implicit source of knowledge: if there is no explicit source named, knowledge can implicitly still be attributed to the author ('*these results suggest…*') or an external source ('*It is generally believed that...*')
- No source of knowledge: the source of knowledge can be absent entirely, e.g. in factual statements, such as '*transcription factors are the final common pathway driving differentiation*'.

Table 1 summarizes our proposed classification.

3 Epistemic evaluation markers

To use our taxonomy to find instances and classes of epistemic evaluation in text, we need to know with what lexicogrammatical cues they are typically marked. Table A1 in the Appendix shows the details, but in summary, a literature review shows widespread agreement on the following cue types:

- Modal auxiliary verbs (e.g. *can, could, might*)
- Qualifying adverbs and adjectives (e.g. *interestingly, possibly, likely, potential, somewhat, slightly, powerful, unknown, undefined*)
- References, either external (e.g. *'[Voorhoeve et al., 2006]'*) or internal (e.g. '*See fig. 2a*').
- Reporting verbs (e.g. *suggest, imply, indicate, show, seem* - see. e.g. Thomas and Hawes (1994) and Hyland (2005) for examples and definitions)

We decided not to add two further categories of epistemic evaluation cues that are often mentioned:

Personal pronouns. ('*we*', '*our results*', or similar). Closer analysis of the papers that mention this shows that in all cases where personal pronouns are mentioned as a hedging device, epistemic verbs are present, in phrases such as: '*we show'*, '*our results suggest'*, etc. Therefore, simply mentioning personal pronouns does not add a useful feature; it does lead to a great deal of false positives, since (first-)personal pronouns are often used in describing methods ('*next, we injected'*, etc.)

Concept	Values
Value	0 - Lack of knowledge
	1 – Hypothetical: low certainty
	2 – Dubitative: higher likelihood but short of complete certainty
	3 – Doxastic: complete certainty, reflecting an accepted, known and/or proven fact.
Basis	R – Reasoning (*'Therefore, one can argue…'*)
	D – Data (*'These results suggest…'*)
	0 – Unidentified (*'Studies report that…'*)
Source	A - Author: Explicit mention of author/speaker or current paper as source *('We hypothesize that…'; 'Figure 2a shows that…'*)
	N - Named external source, either explicitly or as a reference ('*…several reports have documented this expression [11-16,42].*')
	IA - Implicit attribution to the author ('*Electrophoretic mobility shift analysis revealed that…*')
	NN – Nameless external source ('*no eosinophil-specific transcription factors have been reported…*')
	0 – No source of knowledge ('*transcription factors are the final common pathway driving differentiation'*)

Table 1: Proposed classification for epistemic modality and knowledge attribution

In a similar vein, passives are sometimes suggested as an indication of epistemic evaluation, but since they are e.g. often used in Methods sections ('*the rats were injected…*') they do not indicate markers of epistemic modality or attribution.

4 Small Test of Correlation between Epistemic Types and Cues

Using these four features, we want to explore whether all cases where epistemic evaluation occurs are covered by these cues; conversely, do the unmarked cases not have any cues? In other words, are the cues any good at identifying epistemic evaluation, and do certain clues identify certain types?

To investigate these issues, we conducted a small corpus study on two full-text papers in biology (Voorhoeve et. al, 2006; Zimmermann et al., 2005). First, we manually parsed them into clauses via the criteria outlined in (de Waard and

Pander Maat, 2009), leading to a total of 812 clauses. For each clause, we identified the epistemic/knowledge attribution value/source/basis according to the taxonomy in Table 1. Next, we identified the incidence of the four cue types under investigation: modal auxiliary verbs, qualifying adverbs/adjectives, reporting verbs (clauses containing a reporting verb and subordinate clauses controlled by matrix clause with a reporting verb), and references. A sample of this markup, with the clause, attribution/evaluation type, and presence or absence of markers, is given in Table A1.

This sample is too small to draw any quantitative conclusions from. However, we do believe our results support the validity of our model, in two ways: first, because we easily can identify a modality type (value/source/basis) for each of the 812 clauses, and second, because all statements of value < 3 are indicated by one of the four cue types which we have identified.

Next to these general findings, a few correlations between cue type and epistemic evaluation type become apparent (for details, see Table A2):

- Modal auxiliary verbs ('*might, can, could'*) mark potentiality; in our sample, they only indicate clauses of 'possible' value (=1).
- Lack of cues indicates certainty. 47 out of 144 segments with value = 3 have no epistemic cues and no segments of value < 3 have no cues.
- Validating adverbs and adjectives rarely occur; when they do, they usually refer to 'Certain' segments (value = 3). These indicate focus and aim to draw attention to a finding or statement, and are: *important(ly)* (5x), *interestingly*, *striking* (*example*), *presumably*, and *apparently*.
- References mostly occur in 'Certain' segments. This can be because references usually occur when results are cited (3/D/N) or when reference to a figure is made (3/D/IA).
- Within our corpus, 44 discourse segments could not be classified as containing any type of knowledge attribution or evaluation. These were mostly goal statements ('*To identify this process…*') or methods reports ('*We injected all animals…*'). 16 of these (36%) did have a reporting verb (the reporting verbs used here were *analyze, address, assess, define, determine, identify, investigate*, *localize,* and *test*). 12 of these cases were indeed goal clauses containing a to-infinitive verb form.

These results suggest that a combination of verb tense/aspect as well as semantic verb class should be taken into account when analyzing cues for epistemic modality.

The one epistemic type that remains unidentified is 'lack of knowledge' (indicated by a knowledge value of 0); these are marked by different verb types, not just reporting verbs. These clauses are usually marked by specific negational forms of adverbs, verb forms, or nouns ('*has not been established*', '*is unknown'*, '*yet to be determined*' etc. – see Table 2). Therefore, our markers do not adequately cover the 'lack of knowledge' case and finding these constructions by string matching is probably the best way to automate the identification of open research questions in text.

Overall, however, the most prevalent cue we observe is that of a reporting verb, either directly within a clause or governing it, in a matrix clause construction. Half of all statements with Value = 3, 90% of the statements with Value = 2 and 33% of the statements with Value = 1 either contain or are governed by (i.e. are a subordinate clause to a matrix clause containing) a reporting verb. Since this is such a strongly prevalent marker, we wanted to explore if certain reporting verbs perhaps specifically contribute to a particular type of modality.

In Table 2, we show the reporting verbs vs. the knowledge value found in the 812 clauses that we analyzed. Specifically, particular knowledge values can be associated with certain verbs:

- hypothetical statements are reported with '*hypothesize'* (5 x) and cognitive verbs such as '*think'* and '*suspect'*, though they are also often indicated by a modal auxiliary, as discussed above;
- probable statements are marked by '*indicate'* (12x) and '*suggest'* (18 x);
- statements presumed to be true are indicated by '*find'* and especially '*demonstrate'* (15 x).

Value = 0 (Lack of Knowledge)	establish, (remain to be) elucidated, be (clear/useful), (remain to be) examined/determined, describe, make difficult to infer, report
Value = 1 (Hypothetical)	be important, consider, expect, hypothesize (5x), give insight, raise possibility that, suspect, think
Value = 2 (Dubitative)	appear, believe, implicate (2x), imply, indicate (12x), play a role, represent, suggest (18x), validate (2x)
Value = 3 (Doxastic)	be able/apparent/important /positive/visible, compare (2x), confirm (2x), define, demonstrate (15x), detect (5x), discover, display (3x), eliminate, find (3x), identify (4x), know, need, note (2x), observe (2x), obtain (success/results- 3x), prove to be, refer, report(2x), reveal (3x), see(2x) show (24x), study, view

Table 2: Reporting verbs vs. knowledge value for 2 papers

Since the segments containing these reporting verbs are so pivotal to knowledge attribution, they bear closer scrutiny. Generally these are sentence-initial clauses that adhere to the following word order (where Noun Phrases and Verb Phrases are always present, and the others are optional):

Adverb/Connective + Determiner + Adverb/Adjective + **NP** + Modal + Adjective + **VP** + Preposition

All values found in the 42 clauses of this type in one of the papers we examined (Zimmermann et al. (2005)) are provided in Table 3.

Adverb/ Connective	*thus, therefore, together, recently, in summary*
Determiner/ Pronoun	*it, this, these, we/our*
Adverb/ Adjective	*previous, future, better*
Noun phrase	***data, report, study*; method or reference**
Modal	form of *'to be', will, remain*
Adjective	*often, recently, generally*
Verb	***show, obtain, consider, view, reveal, suggest, hypothesize, indicate, believe***
Preposition	*that, to*

Table 3: Values of Parts-of-Speech for Regulatory segments in Zimmermann (2005)

5 Conclusion and implementations

In summary, we have presented a taxonomy of knowledge assessment and attribution and a set of linguistic cues based on a literature overview of from various fields. A small corpus study indicated that the system is simple to use, yet complex enough to cover the many different ways in which biologists attribute knowledge statements. We find that the majority of cases of epistemic evaluation in biological text is instantiated by regulatory segments governed by a reporting verb, prototypically of the form: '*These results suggest'*.

To see if this correlation to epistemic evaluation holds at larger volumes, we plan to try out the above structure in an NLP environment. To begin this, we are examining the case where Value = 2/3 and Source = (I)A: in other words, the author posits a claim. These clauses constitute a specific subset of Propositional Content, which we are calling 'Claimed Knowledge Updates' (Sándor, Á. and de Waard, A., 2012). We are exploring whether an automated syntactic parsing system, combined with a specific subset of reporting verbs will allow the identification of such authorial claims of new knowledge. We plan to use this knowledge to explore what linguistic changes occur when these Claimed Knowledge Updates are cited, and study how knowledge attribution and epistemic modality erode, in the evolution from a claim to a fact.

Acknowledgments

We wish to thank Eduard Hovy for providing the insight that modality can be thought of like sentiment, and our anonymous reviewers for their constructive comments. Anita de Waard's research is supported by Elsevier Labs and a grant from the Dutch funding organization NWO, under their Casimir Programme.

References

Crompton, P. (1997) Hedging in Academic Writing: Some Theoretical Problems, Eng Spec Purposes, Vol. 16, No. 4, pp. 271-287,1997.

De Haan, F. (1999), Evidentiality and Epistemic Modality: Setting Boundaries. Southwest Journal of Linguistics 18.83-101.

De Waard, A., Pander Maat, H. (2009). Categorizing Epistemic Segment Types in Biology Research

Articles. Wkshp on Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS 2009), September 21-23, 2009.

De Waard, 2012. Anita de Waard, 2012-05-23, "Overview of epistemic evaluation and cues from literature", V1, http://hdl.handle.net/1902.1/18253

Hengeveld, K. & Mackenzie, J. L. (2008), Functional Discourse Grammar: A Typologically-Based Theory of Language Structure. Oxford Univ. Press, 2008.

Hovy, E.H. (2011). Private correspondence.

Hyland, K. (1995). The Author in the Text: Hedging Scientific Writing. Hong Kong Papers In Linguistics And Language Teaching, 18 (1995).

Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. Discourse Studies, Vol 7(2): 173–192.

Kilicoglu H., Bergler S. (2008). Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. BMC Bioinformatics. 2008 Nov 19;9 Suppl 11:S10.

Kim, S-M. Hovy, E.H. (2004). Determining the Sentiment of Opinions. Proceedings of the COLING conference, Geneva, 2004.

Lakoff G. (1972). Hedges: a study in meaning criteria and the logic of fuzzy concepts. Chicago Linguistics Society Papers 1972, 8:183-228.

Latour, B., Woolgar, S. (1979). Laboratory Life: The Social Construction of Scientific Facts. Beverly Hills: Sage Publications. ISBN 0-80-390993-4.

Light M., Qiu X.Y., Srinivasan P. (2004). The language of bioscience: facts, speculations, and statements in between. BioLINK 2004: Linking Biological Literature, Ontologies and Databases 2004:17-24.

Martín-Martín, P. (2008). The Mitigation Of Scientific Claims In Research Papers: A Comparative Study. Int Jnl of English 2008 8(2): 133-152.

Medlock B., Briscoe T. (2007). Weakly supervised learning for hedge classification in scientific literature. ACL 2007:992-999.

Myers, G. (1992). 'In this paper we report': Speech acts scientific facts, Jnl of Pragmatlcs 17 (1992) 295-313

Salager-Meyer, F. (1994), Hedges and Textual Communicative Function in Medical English Written Discourse, English for Specific Purposes, Vol. 13, No. 2, PP. 149-170, 1994.

Sándor, Á. and de Waard, A (2012). Identifying Claimed Knowledge Updates in Biomedical Research Articles, Workshop on Detecting Structure in Scholarly Discourse at ACL 2012 (this workshop).

Tang, H., Tan, S., Cheng, X. (2009), A survey on sentiment detection of reviews, Expert Systems with Applications 36 (2009) 10760–10773.

Thomas, S. and Hawes, Th. P. (1994). Reporting Verbs in Medical Journal Articles, English for Specific Purposes, 1994 13(2), pp. 129-148.

Thompson P., Venturi G., McNaught J, Montemagni S, Ananiadou S. (2008). Categorising modality in biomedical texts.. LREC 2008: Building and Evaluating Resources for Biomedical Text Mining 2008.

Verstraete, J.-C. (2001). Jnl of Pragmatics 33 (2001).

Vincze, V., Szarvas, Farkas, Móra and Csirik, (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes, BMC Bioinformatics 2008, 9 (Suppl 11):S9.

Voorhoeve P.M., le Sage C., et. al (2006). A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Cell*. 2006 Mar 24;124(6):1169-81.

Wilson, T. and Wiebe, J., (2003), Annotating Opinions in the World Press, 2003, SigDAIL

Wilbur W.J., Rzhetsky A, Shatkay H (2006). New directions in biomedical text annotations: definitions, guidelines and corpus construction. BMC Bioinformatics 2006, 7:356.

Zimmermann,N., Colyer, JL, Koch, LE and Rothenberg, ME. (2005). Analysis of the CCR3 promoter reveals a regulatory region in exon 1 that binds GATA-1, BMC Immunology 2005, 6:7-7.

Appendix:

Table A1: Example of markup with epistemic evaluation/knowledge attribution types and markers from Zimmermann (2005) – for table headers see caption.

Clause	Value	Basis	Source	Modal	Adv/ Adj	Refs	RV?	Ruled by RV?
DNase I hypersensitivity indicated that	2	D	IA				1	
a region consistent with exon 1 is active in CCR3 transcription.	2	D	IA					1
Together with our previous data showing that	3	D	A				1	
untranslated exon 1 has an important role in CCR3 transcription [27],	3	D	N		1	1		1
we hypothesized that	1	R	A				1	
nuclear proteins bind to exon 1,	2	D	IA					1
and in turn regulate the transcription of CCR3.	2	D	IA					1
In order to test this hypothesis,							1	
a double-stranded oligonucleotide probe that corresponds to bp +10 to +60 of the CCR3 gene was prepared,	3	0	NN					
referred to as E1-FL (exon 1- full length, Figure 2A).	3	D	A			1	1	
This is the exact sequence	3	D	N					
that was deleted in the CCR3(-exon1).pGL3 plasmid	3	D	N					
that demonstrated decreased activity	3	D	N				1	
compared to the full length 1.6 kb construct [27].	3	D	N			1	1	1
Nuclear extracts from AML14.3D10 cells were incubated with the probe								
and resolved on a polyacrylamide gel.								
Two bands were visible (Figure 2B).	3	D	IA			1	1	
The upper band was eliminated	3	D	IA				1	
when 150x molar excess of the unlabelled probe was used (CC: E1-FL in Figure Figure2B),2B),						1		
indicating that	2	D	IA				1	
this is the specific band.	2	D	IA					1
The specific band was eliminated with E1-B and E1-C cold competitors	3	D	IA				1	
indicating that	2	D	IA				1	
the factor binds in the region between +25 and +60 (Figure 2B).	2	D	IA			1		1
In summary, these data indicate	2	D	IA				1	
the presence of proteins in the nuclei of AML14.3D10 cells that bind to CCR3 exon 1 between bp 25 and 60.	2	D	IA					1

'Modal' = containing a modal auxiliary verb; 'Refs' = containing a reference; 'Adverb/Adj' = containing a qualifying adverb or adjective; 'RV' = Reporting verb; 'Ruled by RV' = in a subclause ruled by a matrix clause containing a reporting verb.

Table A2: Correlation between modality type (rows) and modality cues (columns) for two full-text papers

Value	Basis	Source	Modal Aux	Reporting Verb	Ruled by RV	Adverbs/ Adjectives	References	None	*Total*
3	0	0						8	*8*
3	0	IA		5	2	2			*9*
3	0	N		8	5	2	8	2	*25*
3	0	NN	1	2	2			12	*17*
3	D	A		20	1		16	2	*39*
3	D	IA		33	6	1	9	17	*62*
3	D	N		7	7	1	8	6	*29*
3	D	NN		3					*3*
3	R	IA		2	1	1			*4*
3	R	NN		1					*1*
Total value = 3			*1 (0.5%)*	*81 (40%)*	*24 (12%)*	*7 (4%)*	*41 (20%)*	*47 (24%)*	*201(100%)*
2	0	N			1		1		*2*
2	0	NN		1	1				*2*
2	D	0			1				*1*
2	D	A		1					*1*
2	D	IA		22	17		1		*40*
2	D	NN		1					*1*
2	R	0			2	1	1		*4*
2	R	IA		2			1		*3*
2	R	N		1	1				*2*
2	R	NN		1					*1*
Total Value = 2			*0*	*29 (51%)*	*23 (40%)*	*1 (2%)*	*4(7%)*	*0*	*57(100%)*
1	0	0			1				*1*
1	0	NN	1	1	1		1		*4*
1	D	IA	5	5	3	1			*14*
1	R	A	2	2	5				*9*
1	R	IA	1	1					*2*
1	R	NN		2	1				*3*
Total Value = 1			*9(27%)*	*11(33%)*	*11(33%)*	*1(3%)*	*1(3%)*	*0*	*33(100%)*
0	0	0		6	1				*7*
0	0	N		1			1		*2*
0	D	0			1				*1*
0	D	N			1				*1*
0	D	NN				1			*1*
0	R	A		1					*1*
0	R	IA		1					*1*
Total Value = 0			*0*	*9 (64%)*	*3 (21%)*	*1(7%)*	*1(7%)*	*0*	*14(100%)*
Total No Modality			**0**	**16**	**3**	**0**	**3**	**22**	***44***
Overall Total			*10 (2%)*	*146(23%)*	*64(10%)*	*10(2%)*	*50(8%)*	*69(11%)*	*640(100%)*