A Classification of Research Verbs to Facilitate Discourse Segment Identification in Biological Text

Anita de Waard
Elsevier Labs, Jericho, Vt, USA
UiL-OTS, Universiteit Utrecht
Utrecht, The Netherlands
a.dewaard@elsevier.com

Henk Pander Maat
UiL-OTS, Universiteit Utrecht
Utrecht, The Netherlands
h.l.w.pandermaat@uu.nl

Abstract

To improve the classification of biological texts into rhetorical discourse segments, we need a taxonomy for biological verbs. After reviewing three existing classifications, we created a merged taxonomy that encompasses both biology-specific and scientific discourse-specific elements. We provide a manual classification of 239 unique verbs from two full-text biology papers to this taxonomy, and investigate correspondences with segment type. This leads us to propose a simple model of scientific communication, that might enable further model building.

1 Discourse segment identification in biology

To enable improved access to the argumentational components of scientific text (Buckingham Shum et al. 2007, de Waard et al, 2009), we are using a Discourse Segment Type (Grosz & Sidner, 1986) classification of segments at a clause level. In earlier work we have identified ten discourse segment types, which show to be reasonably exclusive and useful namely (de Waard, 2008; de Waard and Pander Maat, 2010):

- Fact: a claim that has been accepted to be true, a known fact.
- Problem: unresolved, contradictory, or unclear issue
- Hypothesis: a proposed idea, not supported by evidence
- Goal: the research goal
- Method: experimental method or protocol
- Result: the outcome of an experiment
- Implication: an interpretation of the experimental results
- Regulatory segments: clauses introducing other clauses (often matrix clauses, e.g. 'These results show that…')
- Intratextual segments: referring text within the document
- Intertextual segments: referring to other documents

Since our segment types are identified at the clause level (simply put, every segment has a verb, for details see de Waard and Pander Maat, 2010) this method provides a more granular identification of argumentational elements compared to other levels of argumentational identification, such as Argumentative Zones (Teufel, Carlotta and Moens, 1999), BioEvents (Nawaz, Thomson et al., 2010), and CoreSC (Liakata, 2010).

To enable the automated identification of these discourse segment types, we are investigating the use of three lexicogrammatical properties, one of which we will discuss here. Work on verb tense/mood/voice (preliminary results and interpretations is provided in de Waard and Pander Maat, 2010); modality (specified through a set of modality markers; in progress), is a work in progress; and verb class: the semantic category that a verb belongs to which is the topic of this paper.

2 Proposed Taxonomy

For our work, we identified 239 distinct verbs from two full-text biology papers, (Voorhoeve et al., 2006) and (Louiseau and Millan, 2009). We attempted a classification of these verbs through six verb classification schemas, three of which have their basis in linguistics:

- Biber' syntactic verb classification (e.g., Biber and Jones, 2005)
- Verbnet (e.g., Kipper et al. 2000), based on Levin's classification on the basis of diathesis alternations (Levin, 1993).
- Korhonen and Briscoe (2004), where an extensive classification effort was undertaken by four human experts for a Gold Standard of classes for biology verbs.

The other three verb classifications we investigated were developed within genre studies, and

focus on the rhetorical goal of authors in specific textual contexts:

- Thomas and Ye (1991) who identify textual, mental and research verbs,
- Thomas and Hawes (1994) who classify the reporting verbs used in medical journal text into Discourse Verbs, Real-World or Experimental Verbs and Cognition Verbs,
- Williams (1996) studies the correspondence between verb class, verb form and article section and defines seven categories: reporting, observation, relations, defining cause and effect, change and growth, and methods.

However, none of these completely suit our needs. The Levin classes and associated VerbNet classes are too fine-grained and did not contain enough verbs of the corpus-specific verbs that we needed to classify. For example, VerbNet only has gives four verbs the type 'Implicate', out of 28 forms that we find. Likewise, in the biology-specific classifications determined by Korhonen and Briscoe (2004), many of our verbs can not be classified, whereas it contains many verbs that our texts did not use. This is probably due to the highly specified nature of biological text, and the fact that their classification was specific to a particular area of biology. On the other hand, the reporting-verb focused approach in genre studies such as Thomas and Hawes is more concerned with relations between authors than those between proteins, and ignores technical verbs altogether.

We have therefore made a taxonomy that combines these different approaches. The taxonomy, together with a list of verbs for the two full-text biology papers classified to this taxonomy, and comparable terms in the other classifications are given in the Appendix. Several verbs fit in more than one category, which is unhelpful if we wish to use the categories for computational identification of segment types. In general, a disambiguation can be done using linguistic context, e.g. 'remain to be seen' is a typical context for a Cognition version of the word 'remain', as opposed to 'remains constant' which implies Change and Growth.

3 Results

The results of a cross-correlation of verb class with segment type for the two annotated full-text articles is given in Table 1. Four clusters of verb class to segment types can be seen:

- Discourse verbs are most often used for Regulatory and Inter- and Intratextual segments;
- Second, there is a cluster of segments that

	Fact	Problem	Hypothesis	Goal	Method	Result	Implication	Regulatory	Intertextual	Intratextual	*Total*
Discourse Verbs	0	2	0	2	1	1	0	**7**	**4**	**12**	*25*
Investigation	1	9	0	**31**	**13**	**4**	2	3	0	0	*63*
Procedure	1	2	0	0	**126**	**8**	1	0	2	0	*140*
Observation	0	1	1	0	**3**	**29**	2	5	0	0	*41*
Total Research Verbs	*2*	*12*	*1*	***31***	***142***	***41***	*5*	*8*	*2*	*0*	*244*
Prediction	0	1	6	0	0	0	**1**	**11**	0	0	*19*
Interpretation	1	2	2	1	3	4	**31**	**38**	0	0	*82*
Comparison	0	0	1	1	1	0	**7**	**6**	0	9	*25*
Cognition	0	7	0	1	3	0	**1**	**5**	0	0	*17*
Total Sensemaking Verbs	*1*	*10*	*9*	*3*	*7*	*4*	***40***	***60***	*0*	*9*	*143*
Cause, Effect	**18**	**5**	**27**	2	1	**79**	**55**	0	0	0	*187*
Change and Growth	**4**	**0**	**0**	0	1	**21**	**4**	0	0	0	*30*
Properties	**19**	**2**	**6**	0	2	**42**	**8**	1	0	1	*81*
Total Properties and Relations	**41**	**7**	**33**	2	4	**142**	**67**	1	0	1	*298*
Total segment type	*44*	*31*	*43*	*38*	*154*	*188*	*112*	*76*	*6*	*22*	*710*

Table 1: Verb class vs. Segment Type, manual classification for (Voorhoeve et. al, 2006) and (Louiseau and Millan, 2009). Totals and subtotals are in italics; clusters mentioned in the text are given in bold.

pertain to the experimental activities that make up a paper: Goal/Method/Results. We see a strong correlation with Research verbs here; a special category is formed by the Procedural verbs, which are used, in the majority of the cases, for Methods segments.

- Third, there is a correlation between Implication and Regulatory segments and sensemaking verbs. Clearly, sensemaking takes place between statements pertaining to either research or objects of study in these 'transitional' fragments.
- Fourth, there is a correlation between Properties and Relations and two types of segments: either Facts of Hypothesis statements (known or postulated facts about the world) and experimental Results and Implications.

4 Discussion

4.1 Interpretation, based on a simple model of scientific research

We can interpret these results by offering a simple model of experimental science. In our simplistic model, scientific discourse consists of the interaction of three types of entities: Concepts, (real-world) Objects, and People (scientists). By Concepts, we mean entities which are not tangible, such as processes ('oncogenisis'), theoretical concepts ('downregulation'), or categories of entities grouped by function ('beta-blockers'). By Objects we mean items that are tangible, and have a representation in the physical world, such as cells, creatures, dials on measuring devices, etc. We immediately concur that this distinction is not always clear-cut, and the vagueness between these two entity types is one of the factors that complicates this simplistic conceptual view of science communication.

The separation of people from the other two is quite clear: humans are the ones who plan and perform experiments to and debate the other entities. If we now try to explain the prevalent use of verbs with segment types, the model ensues:

- Research verbs describe human activities pertaining to Objects, either preparing them or observing their behavior.
- Discourse verbs have to do with interpersonal communication, including argumenttion, cognitive activities, and the formation and interpretation of Concepts.
- Sensemaking verbs pertain to argumentation that takes the observations of Objects and projects them into the Conceptual space, as models are discussed or created, and findings interpreted.
- Properties and Relations can pertain to Concept-Concept or Object-Object relations.

Figure 1 shows a sketch of how, in this world model, Verb Class corresponds to a relationship between various entity types.

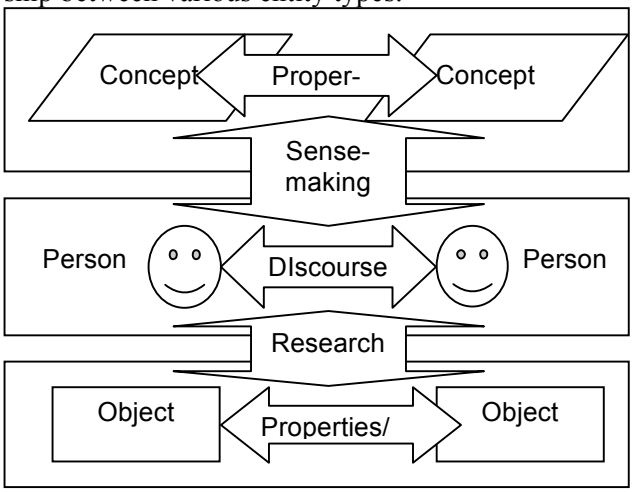

Figure 1: A simplistic view of scientific discourse, related to our verb class taxonomy

4.2 Conclusion

In conclusion, we find that our verb class taxonomy is fairly useful to identify a specific class of research activities, through the identification of segment type. Together with verb form and modality, we hope to use this classification to be able to semantic segment types of clauses in biological text, in able to achieve access to the rhetorical structures underlying this discourse. We are seeking partnerships with computational linguists to investigate this in a more quantitative and scalable way.

Acknowledgements

We are grateful for the comments from the workshop reviewers, and the very helpful suggestions from Ellen Hays.

References

Biber, D. and Jones, J.K., (2005). Merging corpus linguistic and discourse analytic research goals, *Corpus Linguistics and Linguistic Theory* 1(2) (2005), 151_182

Buckingham Shum, S.J., Uren, V., Li, G., Sereno, B. and Mancini, C. (2007). Modeling Naturalistic Argumentation in Research Literatures. Int. *Jnl. of Int. Systems, Spec. Issue on Comp. Models of Natural Argument, Eds: C. Reed and F. Grasso*, 22, (1), pp.17-47

De Waard, A. and Pandermaat, H. (2010), Categorizing Epistemic Segment Types in Biology Research Articles. To be published as a chapter in *Linguistic and Psycholinguistic Approaches to Text Structuring*, Laure Sarda, Shirley Carter Thomas & Benjamin Fagard (eds), John Benjamins, (planned for 2011).

De Waard, A., Buckingham Shum, S.J., Carusi, A., Park, J., Samwald, M. and Sándor, Á.. (2009), Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims, In: *Proc. of the Wkshop on Semantic Web Applications in Scientific Discourse (SWASD 2009),* co-located with ISWC-2009.

De Waard, A., A Pragmatic Structure for the Research Article (2007). In: *Proceedings ICPW'07: 2nd International Conference on the Pragmatic Web, 22-23 Oct. 2007, Tilburg: NL*. (Eds.) Buckingham Shum, S., Lind, M. and Weigand, H. Published in: ACM Digital Library & Open University ePrint 9275.

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12: 175-204.

Kipper, K., Dang, H. T., & Palmer, M. (2000). Class-based construction of a verb lexicon. In: *AAAI/IAAI*. pp. 691–696.

Korhonen, A. & Briscoe, T. (2004). Extended lexical-semantic classification of English verbs. In: *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*. Boston, MA.

Liakata, M. (2010) Zones of conceptualisation in scientific papers: a window to negative and-speculative statements, Roser Morante and Caroline Sporleder (eds.) *Proc. of the Workshop on Negation and Speculation in Natural Language Processing*, (NeSp-NLP 2010), July 2010

Loiseau, F., Millan, M.J. (2009). Blockade Of Dopamine D3 Receptors In Frontal Cortex, But Not In Sub-Cortical Structures, Enhances Social Recognition In Rats. *European Neuropsychopharmacology* 2009,19 (1) 23-33.

Levin, B. (1993). *English verb classes and alternation, A preliminary investigation*. The University of Chicago Press, 1993.

Nawaz, R., Thompson, P., McNaught, J. Ananiadou, S. (2010). Meta-Knowledge Annotation of Bio-Events. *In Proceedings of LREC* 2010, pages 2498-2505.

Teufel, S., J. Carletta and Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles, In: *Proc EACL* 1999.

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

Thomson, G. and Ye, Y. (1991). Evaluation in the Reporting Verbs Used in Academic Papers, *Applied Linguistics* 12: 365-382 (1991).

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.

Williams, I.A. (1996). A Contextual Study of Lexical Verbs in Two Types of Medical Research Report: Clinical and Experimental. *English for Special Purposes*, Vol. 15, No, 3, spp. 175-197, 1996

Class/Subclass	Unique verb occurrences in two biology papers.
Discourse verbs *Thompson and Ye: 'Textual verbs'*	address, base, depict, describe, mention, note, report, represent, review, **show**, study, **suggest**, term
Research Verbs *Thompson and Ye: 'Research verbs'*	
Investigation *Thomas and Hawes: 'Procedural verbs';* *Williams: 'Methods'*	**compare**, **demonstrate**, **detect, determine**, **elucidate**, **evaluate**, **examine**, **exclude**, exemplify, expose, extend, find, **identify**, investigate, pinpoint, mimic, **remain**, **require**, require, shed [light], start identify, strengthen, **substantiate**, test, verify
Procedure *Thomas and Hawes: 'Procedural verbs';* *Williams: 'Methods'*	accumulate, **activate**, adapt, administer, affix, allow recover, analyze, anesthetize, annotate, base, calculate, characterize, clone, **compare**, conduct, conform, contain, connect, conserve, consist, construct, **control**, cotransfect, correspond, create, derive, **determine**, develop, dissolve, divide, drill, employ, enrich, **evaluate**, **express**, find, follow, frozen, generate, handle, harbour, house, immortalize, **impair**, implant, include, infuse, insert, introduce, involve, keep, leave, localize, look, lose, lower, make, minimize, mix, model, mount, mutate, obtain, overcome, overlap, perform, permit, place, possess, present, prevent, purchase, reduce, remove, replace, resemble restrain, retract, section, serve, share, spend, stabilize, synthesize, take, transduce, transfect, **use**
Observation *Thomas and Hawes: 'Objective verbs'* *Williams: 'Observation verbs'*	**characterize**, **compare**, **correlate**, detect, detect, **express**, **find**, **identify**, monitored, note, observe, see, **show**
Sensemaking Verbs	
Prediction *Thomas and Hawes: 'Pre-experiment verbs'*	**elucidate**, hypothesize, **involve**, point to, predict, propose, provide [indication], raise, **remain**, seem, **suggest**
Interpretation *Thomas and Hawes: 'Post-experiment verbs'*	associate, conclude, conjecture, **demonstrate**, **exclude**, explain, **implicate**, indicate, provide, provide [evidence], reveal, **show**, stress, **substantiate**, **suggest**, support, underpin
Comparison *Hyland: 'Evaluative verbs'*	**compare**, confirm, expect, provide, underpin, validate
Cognition *Thomas and Hawes: 'Cognition verbs'/Biber: 'Mental verbs'*	choose, concern, decide, emphasize, **examine**, **exclude**, infer, judge, know, **remain**, take [advantage of]
Object Properties and Relations	
Cause and Effect *Williams: 'Cause and Effect'*	abolish, abrogate, accelerate, act, affect, allow, attenuate, block, bypass, cancel, cause, circumvent, collaborate, confer, connect, contribute, **control, correlate**, degrade, depend, disinhibit, disrupt, encode, enhance, exert, **express**, facilitate, fail [to express], fail [to discriminate], have [an effect], **impair**, **implicate**, improve, induce, inhibit, **involve**, lead, make [resistant to], mediate, modify, neutralize, numb, obtain, **overcome**, participate, permit, play [a role], predict, prevent, provoke, reduce, reflect, regulate, reinforce, relate, replace, **require**, result, reverse, **show**, silence, stimulate, suppress, target, undergo, underlie, **use**, yield
Change and Growth *Williams: 'Change and Growth'*	amplify, cease, continue, disrupt, downregulate, exert, expand, **express**, grow, increase, mimick, proliferate, reach, **remain**, **show**, **spend**
Properties *Williams: 'Defining verbs' (is a subset)*	accumulate, **activate**, **characterize**, conform, conserve, consist, contain, correspond, divide, enrich, exist, express, **find**, harbour, have, **impair**, include, **involve**, localize, lose, **overcome**, overlap, possess, resemble, share, **spend**, stabilize

Appendix. Instances of the verb class taxonomy for two full-text biology papers Voorhoeve and Louiseau).
On the left, our taxonomy; in italics, overlap with other taxonomies. On the right, verbs found in two full-text biology papers, classified according to this taxonomy. Bold indicates that verbs occur in more than one category.