BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology

by

Andrew Eric Dolbey

B.A. (Towson State University) 1992
M.A. (University of California, Berkeley) 1995

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Linguistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Charles J. Fillmore, Chair
Professor Jerry Feldman
Professor Sharon Inkelas
Professor Martha Palmer

Fall 2009

BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology

Andrew Eric Dolbey

Abstract

BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology

by

Andrew Eric Dolbey

Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor Charles J. Fillmore, Chair

In this study I introduce BioFrameNet, an extension of the Berkeley FrameNet lexical database to the domain of molecular biology. I examine the syntactic and semantic combinatorial possibilities exhibited in the lexical items used in this domain in order to get a better understanding of the grammatical properties of the language used in scientific writings on molecular biology.

The particular data considered is a collection of Gene References in Function (GRIF) texts that describe various types of intracellular protein transport events, a collection that had previously been annotated for an ontologically grounded knowledge base. GRIF texts use long, complex noun phrases, with the omission of many items, resulting in a dense, telegraphic style of writing. This introduces an additional level of complexity to language used in scientific writings of this domain.

In providing a frame semantic analysis and cataloging of the grammatical structures used in the scientific language of molecular biology, we see how well a FrameNet approach can handle language of this domain. Extending FrameNet to this domain serves as a testing ground for some of FrameNet's principles and claims, as it becomes evident how well a FrameNet approach handles language in a significantly different field than has been previously examined. I show how domain ontologies and knowledge bases, sources of definitions and classifications of biological phenomena based entirely on their biological properties, can be used in conjunction with lexical resources. At the same time, I also illustrate the overlap of grammatical properties across separate domain ontology classes, demonstrating that although the biology defined and classified in these classes is different, language used to describe and discuss them is not. Finally, I also explore the possibility that BioFrameNet can be used with tools that carry out Natural Language Processing tasks such as automatic semantic role labeling. Therefore, this work is at the intersection of theoretical frame semantics and practical applications and will potentially provide benefit to linguists, BioNLP engineers, and biologists.

Table of Contents

Acknowledgements

First and foremost, thank you to my dissertation committee members - Charles Fillmore, Jerry Feldman, Sharon Inkelas, and Martha Palmer. While none of these individuals' specialty is molecular biology, each one of them was willing to learn enough about it to understand the project and provide significant input and response. And tusen takk to Pastor Chuck who was willing to be a part of my dissertation work even though he had officially retired several years before. It was an honor to have him as the chair of the committee.

During the time I spent in the bioinformatics lab run by Larry Hunter at the University of Colorado Health Sciences Center, School of Medicine, I learned a tremendous amount about molecular biology and the language used to talk about topics in this field. It is from this group that I obtained the data on which BioFrameNet is based. I am deeply indebted to various members of this group, especially Philip Ogren, Mike Bada, Kevin Cohen, Helen Johnson, and of course, Larry Hunter.

Thank you to the University of Colorado Health Sciences Center and the University of California Berkeley for the various forms of financial support provided over the span of this project.

Special thanks to The MITRE Corporation for their flexibility and support, and for allowing me to complete this dissertation while working for them full time. Statement of this affiliation is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed in this dissertation.

Thank you to the Fulbright Program and The American Scandinavian Foundation for funding my two-year visit in Finland. The environment of the Linguistics Department in Helsinki was my first extended exposure to computational linguistics, which in many ways changed the direction of my research, and career. Jan-Ola Östman was a tremendous source of support and friendship during this visit.

Huge thanks to Belen Flores, graduate assistant in the Linguistics Department at UC Berkeley who has supported me since I first started there in 1993.

Thank you to fellow grad students – Gunnar Hansson, Matthew Juge, Tess Wood, Natasha Warner, Ju Namkung, Bill Weigel, Ron Sprouse, and Dave Peterson.
Sincere thank you to fellow FrameNetters Collin Baker, Michael Ellsworth and Miriam Petruck. Your contributions and support have been invaluable.

Thank you to my family who may not always have understood what I was doing, or why, but still supported me all these years.

Thank you to some of my undergraduate professors at Towson State University who were so inspiring - Carl Behm, Jenny Jochens, Fran Rothstein, and Barbara Leons.

Special thanks to American Field Service which sent me to Norway as a 16-year old high school exchange student. Warmest gratitude to my host family there, Familien Håland, where my life as a linguist started as I struggled to learn Jærsk, my first second language! And particular thanks to Magnhild, my then 5-year old host sister who tortured me endlessly and forced me to learn this dialect.

Most importantly, I owe enormous debt and gratitude to my wife, Andrea Huseth. She has provided care, support, and interest throughout the duration of the project. And now I get to return to life with her.

Chapter 1

Introduction

1.1 BioFrameNet: a FrameNet extension to domain of molecular biology

Over the last two decades there has been a dramatic increase in the frequency of publication of scientific texts on molecular biology, as reported in Cohen and Hunter (2005). Thousands of new publications are recorded per week in Medline, the National Library of Medicine's (NLM) primary bibliographic database, across thousands of different scientific journals.[1] This increase is tied to recent advances in rapid, high throughput genomics technology, in which analyses of all the genes of an entire genome can be carried out in a matter of hours. Results of such experiments are most often reported in journal articles. What's more, new databases that have been created and expanded to store these results are often directly linked with the literature, and sometimes hold their own pieces of scientific text about the results.

In order to take advantage of this volume of literature, there has been much effort expended to build tools for automatic processing of molecular biology texts. A common goal is to enable extraction of facts and relationships asserted and described in them. Resources typically required for building such tools, such as named entity (NE) identifiers and part of speech (POS) taggers, must often be adapted for handling domain specific texts. For example, POS taggers might require lexical items not included in general language versions. Tasks like extraction of complex assertions from texts require capabilities in higher-level language processing, in particular the ability to link syntactic and semantic elements of lexical units, phrases, and clauses. FrameNet (Ruppenhofer et al. 2006) and PropBank (Kingsbury and Palmer 2002) are two well known lexical resources aimed at providing descriptions of these sorts of linkings for general English. Higher level resources like these might also need to be adapted when targeting specific domains.

In this study I will introduce BioFrameNet, an extension of the Berkeley FrameNet lexical database to the domain of molecular biology.[2] I will examine the syntactic and semantic combinatorial possibilities exhibited in the lexical items used in this domain in order to get a better understanding of the grammatical properties of the language used in scientific writings on molecular biology. In providing a frame semantic analysis and cataloging of the grammatical structures used in the scientific language of molecular biology, language with different sorts of complexities, we will see how well a FrameNet approach can handle language of this domain. Extending FrameNet to this domain will serve as a testing ground for some of FrameNet's principles and claims, as it becomes

[1] See NLM Fact Sheet (http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html) for further discussion of the «explosive growth» of biological information.

[2] For more information about FrameNet, see the information page on the project's website: http://framenet.icsi.berkeley.edu/index.php?option=com_content&task=view&id=40&Itemid=1

evident how well a FrameNet approach handles language in a significantly different field than has been previously examined. The extension will follow FrameNet's grounding in frame semantics (Fillmore 1982, 1985; Fillmore and Atkins 1992; Petruck 1996; Gawron 2008), and in doing so will provide a new perspective on the language of this domain. I will show how domain ontologies and knowledge bases, sources of definitions and classifications of biological phenomena based entirely on their biological properties, can be used in conjunction with lexical resources. At the same time, I will also illustrate the overlap of grammatical properties across separate domain ontology classes, demonstrating that although the biology defined and classified in these classes is different, language used to describe and discuss them is not. Using the criteria for Frame creation and separation proposed by the FrameNet team (Ruppenhofer et al. 2006), I will argue that creating separate Frames for these classes is not warranted. Instead, the overlapping cases will be collapsed into a smaller set of Frames. Finally, I will also explore the possibility that BioFrameNet can be used with SemLink, an approach for linking together different resources for use in a tool built for automatic semantic role labeling (ASRL). Therefore, this work is at the intersection of theoretical frame semantics and practical applications and will potentially provide benefit to linguists, BioNLP engineers[3], and biologists.

This chapter outlines the background, goals, and methods of the rest of the dissertation. In section 1.2 I describe the frame semantic approach adopted in this study, and present a description of scenarios typical of this domain. Section 1.3 presents the particular type of intracellular event that is the focus of BioFrameNet, namely *protein transport*. The corpus data analyzed for the Frames proposed in this study are presented in section 1.4. In 1.5 I illustrate ontological analysis of this corpus data in a domain knowledge base, and discuss differences between this analysis and that of BioFrameNet. Section 1.6 presents an example of a natural language processing (NLP) tool making use of the grammatical and frame semantic information provided by BioFrameNet. Finally, in 1.7 I outline the organization of the rest of the dissertation.

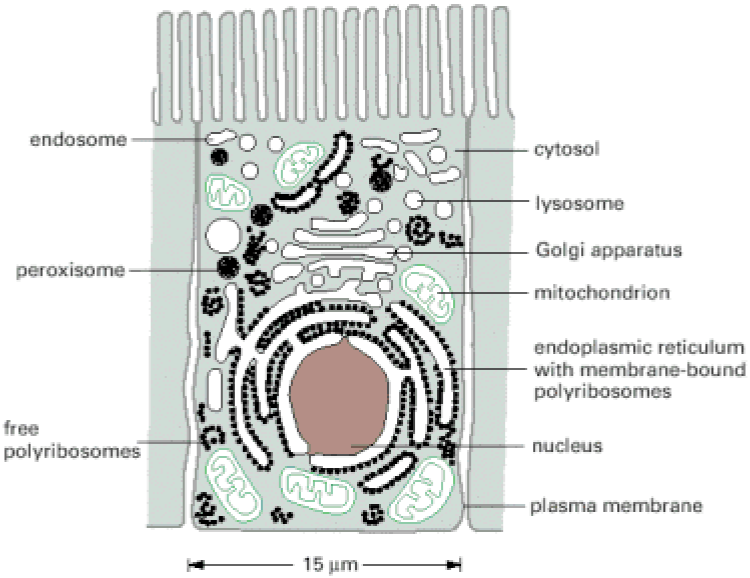
1.2 Onomasiological approach of FrameNet and BioFrameNet

As will be discussed in the next chapter, the primary unit of analysis in FrameNet is the "Frame", a conceptual structure that describes a situation, object, or event along with participants and props associated with it. From an onomasiological perspective (Koch 2008; Geeraerts 2006; Grondelaers and Geeraerts 2003), the scenarios captured in Frames are starting points. The goal is to 'encode' scenarios, and show how they can be described linguistically with lexical units (LUs) that evoke them and grammatical structures that provide details about the participants and props of the scenario. BioFrameNet shares this approach; it holds molecular biology scenarios as starting points and attempts to show how they are encoded with particular lexical units and grammatical structures.

[3] 'BioNLP' is the accepted abbreviation for natural language processing of biomedical texts.

The most important things to know about for understanding texts considered in this study are biological events that take place within and outside cells, and entities that participate in such events.[4] The texts considered here discuss eukaryotic cells, cells which contain more than a dozen membrane-bound subcellular components where specific activities take place. The membrane that surrounds these components is a layer that allows only certain things to pass through it.[5] The most important of these components is the **nucleus**, a compartment that holds the cell's DNA. Other important components include the **endoplasmic reticulum** and the **Golgi apparatus**, structures that are responsible for production and processing of proteins. Surrounding the cell is the plasma membrane, a structure similar to those of the membranes that surround subcellular components. The region of the cell between the nucleus and the plasma membrane is the **cytoplasm**. Figure 1 shows a diagram that illustrates these items.

(1) Structure of eukaryotic cell (taken from Alberts et al. 2002, Chapter 12, Figure 1)



The major intracellular compartments of an animal cell. The cytosol *(gray),* endoplasmic reticulum, Golgi apparatus, nucleus, mitochondrion, endosome, lysosome, and peroxisome are distinct compartments isolated from the rest of the cell by at least one selectively permeable membrane. (Alberts et al. 2002, Chapter 12, Figure 1)

Events and processes that happen in cells include: **metabolism**, a process that occurs in two types, the breaking down of molecules to produce energy and the constructing of complex molecules used for other functions; **cell signaling**, a communication system in which cells respond to signals in their external environment, a primary means of regulating internal activities; **cell division**, the creation of new cells, a process in eukaryotic cells called *mitosis*; and **protein synthesis**, a complex process that includes

[4] This brief description is based on a primer provided by NLM, and a wikipedia entry on cell biology. http://www.ncbi.nlm.nih.gov/About/primer/index.html, http://www.ncbi.nlm.nih.gov/About/primer/genetics_cell.html; cellular and molecular biology texts books like Alberts et al. (2002) and Lodish et al. (1999) provide much more detailed information.

[5] 'Selectively permeable bilayer' is the technical description of this sort of structure.

the generation of messenger RNA (mRNA) from a portion of the DNA in the nucleus, the movement of the mRNA out of the nucleus into the cytoplasm, and lastly, the creation of proteins for modulation and maintenance of cellular activities.

The language used for naming these events and the molecular entities that are their participants includes much domain-specific vocabulary, e.g., items like *glycolysis*, *pyruvate*, *lysosome*, and *peroxisome*. Entity names often include numerals, word-internal punctuation, and chemical notation, e.g., *6-chloro-1,2,3-benzothiadiazole*, *[^{14}C]metformin*, and *3'-PPIn-phosphatase myotubularin 1 (MTM1)*. Abbreviations are typically provided for these names, though frequently, as in the last example just listed, the mapping between the full name and the abbreviation is not obvious. Another very typical linguistic phenomenon seen in language of this domain is long compounds, e.g., *human equilibrative nucleoside transporter 3*. Abbreviations for compounds like this are usually concatenations of the first letters of each word, e.g., *hENT3* for this case.[6]

Molecular biology jargon and entity names often would not be understood by non-specialists. What's more, general English is frequently used with special meanings. For example, *transcription* is the term used for the process of creating RNA nucleotide sequences based on DNA sequences, and *translation* is the term for the production of the amino acid chains of proteins based on nucleotide sequences of RNA. These two processes are necessary parts in the creation of proteins, a process referred to as gene *expression*, showing another case of general English language being given a special, domain-specific meaning.

1.3 Event of focus: intracellular protein transport

The type of event most discussed in the texts examined in this study is that of **intracellular protein transport**, the directed motion within a cell of proteins or protein complexes from one place in the cell to another place. Transport of other kinds of structures, e.g., ions, sugars, and other small molecules, also takes place in eukaryotic cells, and is described with similar types of grammatical structures. Nevertheless, the only kind of transport examined in the data source used for this study is that of protein transport. As a result, the focus of this study is also on protein transport.

One of the reasons why proteins are the subject of intensive research is that they are critical participants in all the activities listed in the previous section. And while their structure and chemical composition are crucial properties that enable them to participate in cellular processes, they must be in particular locations within the cell for specific activities to take place successfully. In addition, protein transport phenomena are important because they often form part of the regulation of another process, either enabling and/or increasing the frequency of the other process, or reducing the frequency of and/or terminating the other process.[7] This elaborate orchestration of protein locations

[6] See K. Cohen, A. Dolbey, G. Acquaah-Mensah, and L. Hunter (2002) for discussion of contrast and variability in use of upper and lower case letters in entity names.

[7] After most of the analysis for this dissertation was completed, the Hunter Lab started a project in which they included regulation concepts in the knowledge base they had created, and annotated texts to illustrate how these concepts are expressed in scientific texts.

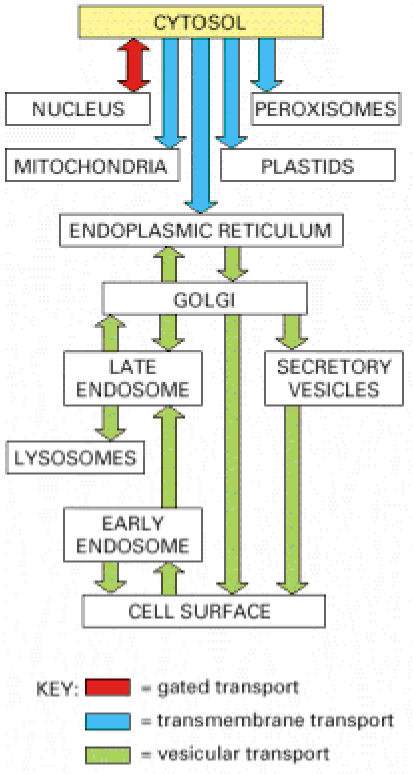
inside a cell is often called *protein sorting*, and is critical for successful functioning of the variety of biological processes associated with specific subcellular components.

There are several types of intracellular protein transport, varying in specific biological mechanisms involved, location within the cell, and types of membranes crossed, if any. In cases of transport across the nuclear membrane, there are control points which allow transport of only certain proteins and in only certain conditions, a process called *gated nuclear transport*. Transport events can also involve crossing membranes with no dedicated control points, a process called *transmembrane transport*. Another mechanism of transport does not involve proteins crossing a membrane, but rather, formation of sacs that hold proteins in one organelle and then fusing with a different organelle. This is known as *vesicular transport*. Figure 2 shows a helpful schema that indicates typical subcellular locations associated with the different types of transport.

The vocabulary of the Frame needed for such concepts includes words like *promote*, *inhibit*, *block*, and *modulation*; Frame Elements include regulated process and regulating entity. The following are two examples of the kinds of texts a regulation corpus would comprise.

> statins inhibited membrane translocation of the small G protein family members Ras and Rho
>
> EBV LMP1 blocks p16INK4 pathway by promoting nuclear export of E2F-4 and E2F-5.

(2) Transport types, locations (taken from Alberts et al. 2002, Chapter 12, Figure 6)



A simplified “roadmap” of protein traffic. Proteins can move from one compartment to another by gated transport *(red),* transmembrane transport *(blue),* or vesicular transport *(green).* The signals that direct a given protein's movement through the system, and thereby determine its eventual location in the cell, are contained in each protein's amino acid sequence. (Alberts et al. 2002, Chapter 12, Figure 6)

Each type of protein transport involves two kinds of elements: locations and proteins. There are two intracellular locations of relevance, the place within the cell where the protein was located before transport, and the place where the protein is located after transport. These are called *transport origin* and *transport destination*, respectively. The protein that moves is of course a critical element of the process. However, occasionally there is also a different protein that initiates or controls transport of the first. In this work, proteins showing one or the other of the two types of roles are called *transported entity* and *transporting entity*, respectively.

The primary goal of this study is the definition of semantic frames for intracellular protein transport. It will be argued that two separate but related Frames are necessary and sufficient for linguistic analysis of texts covered. The critical components of these definitions are lists of predicators that evoke the Frames, frame elements that provide fuller detail about transport events, as described in the previous paragraph, and the relation of these Frames with each other and with other defined semantic Frames. In support of the proposed definitions, a collection of texts annotated with relevant predicators and frame elements will be provided. While the texts annotated also include discussion of other phenomena than protein transport, annotation of these is not systematically included in the data source used for this study, and thus analysis of them is not included in this study.

1.4 Corpus data source

The data for this study come from a collection of texts of a special type called 'Gene References into Function' (GRIF). These are brief statements about the function of a *gene product*, a protein or RNA molecule whose production is based on the biological template a gene provides.[8] GRIFs are included in Entrez Gene[9], a searchable database of genes run by the National Center for Biotechnology Information (NCBI).[10, 11] Each GRIF is associated with two other items, a specific Entrez Gene ID and one or more published papers describing the function.[12] Because GRIFs are designed to be a concise textual description of a gene function, they are limited in length to a maximum of 254 characters. The mandatory limits on length lead to dense, compact language use with frequent noun compounds and other multiword expressions (MWEs), coordination structures, and references to particular domain-specific terminology and named entities (NEs). Also, many times the description of a function in a GRIF makes reference to other functions and processes, often in the form of metonymy. All of these characteristics make the texts of GRIFs more difficult for non-biologists to comprehend, and even for biologists who specialize in a set of genes that does not include the particular gene of the GRIF. Similarly, they add to challenges of automatic processing of GRIFs.

The collection of GRIFs considered here was gathered by a team of researchers in Lawrence Hunter's Bioinformatics Lab at the University of Colorado Health Sciences Center[13], where the team was working to improve techniques and strategies for building a knowledge base for modeling concepts of molecular biology. The particular technique most targeted was that of automated extraction of assertions from molecular biology texts, followed by placement of the concepts denoted by the assertions into the knowledge base. Extracting information from GRIFs for genes with a role in protein transport activities was an initial test case for tools being built for this. In order to facilitate building and testing such tools, the team annotated the full collection of protein transport GRIFs. These annotations covered key concept predicates that were mentioned in the GRIFs, as well as NEs and cellular components, and the role these play in the protein transport event. A few examples of phrases from protein transport GRIFs and their annotations for the knowledge base are provided in (3) below. The annotations use square brackets to separate a portion of text in the GRIF that is a realization of an NE or cellular component; specifications of the particular cellular component or type of NE are

[8] For futher info on GRIFs, see http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html.
[9] Website for Entrez Gene: http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene.
[10] NCBI (http://www.ncbi.nlm.nih.gov/) is a subgroup of the National Library of Medicine (NLM, http://www.nlm.nih.gov/), which is part of the National Institutes of Health (NIH, http://www.nih.gov/).
[11] Although it is possible for any researcher to submit a GRIF for review and possible addition to Entrez Gene, most GRIFs are provided by NLM staff. Staff members responsible for providing GRIFs have advanced degrees in life sciences.
[12] The papers a GRIF is associated with are indicated by an ID for PubMed, a search engine for accessing NLM's database of citations and abstracts of biomedical research articles.
[13] Website for Hunter's Bioinformatics research lab: http://compbio.uchsc.edu/. This lab will be referred to elsewhere in this study as 'Hunter Lab'.

given in subscript notation, followed by a colon and the name of the role the item plays in the protein transport event described by the GRIF.

(3) Portions of protein transport GRIFs

[GLUT4 $_{\text{Protein:TransportedEntity}}$] translocation

[nuclear $_{\text{CellularComponent:Destination}}$] translocation of [endoG $_{\text{Protein:TransportedEntity}}$]

translocation of [p27 $_{\text{Protein:TransportedEntity}}$] from the [nucleus $_{\text{CellularComponent:Origin}}$] to the [cytoplasm $_{\text{CellularComponent:Destination}}$]

In this study, I have borrowed this collection of GRIFs, but have tailored the annotations for a lexical resource in which grammatical structure of the language used to express the concepts is the focus.

In addition to the kinds of structures shown in (3), language used in GRIFs frequently contains complex combinations of verbal and nominal structures. Thorough analysis of such structures will be discussed in detail in this chapter 3.

1.5 Domain ontology classes and BioFrameNet Frames: orthogonal resources

The concepts held in the Hunter Lab protein transport knowledge base are arranged in a hierarchical structure based on biological knowledge of the events and processes being described. The top-level class in this structure is protein transport. Three subclasses, gated nuclear transport, transmembrane transport, and vesicular transport, are defined for the top-level class. For one of these, vesicular transport, the additional subclass of endocytosis has been defined. The overall hierarchy is shown below in figure (4).

(4) HLKB protein transport classes
protein transport
gated nuclear transport
transmembrane transport
vesicular transport
endocytosis

This hierarchy is intended, and designed and created, as a classification of biological phenomena. The slots of these protein transport classes (explained further in Chapter 4) are: transported entity, transporting entity, transport participants (a generalized version of the previous two), transport origin, transport destination, and transport locations (a generalized version of the previous two).

The degree to which the above protein transport ontology corresponds to a classification of the frame semantics observed in language used to discuss these phenomena is one of the primary subjects of analysis in this study. An immediate question that arises in considering the relation between the biological concepts and their language's frame semantics is the number of Frames needed to cover the data in the GRIF collection, and, if more than one, their relation with one another. Following the criteria suggested by Ruppenhofer et al. (2006, pp. 11-19), I will argue that the differences in biology which motivated division of the knowledge base in to the classes listed above are not directly reflected in the grammar and frame semantics of the language used to characterize these differences, and thus do not warrant creating corresponding separate Frames.[14] Yet, at the same time, causative uses of the transport predicates lead to inclusion of agentive frame elements, and thus motivate defining an additional protein transport Frame, for a total of two separate but related Frames.

Definitions of ontology class structure and definitions of linguistic Frame structure offer orthogonal resources which, when carefully combined might offer improved analysis of language data.

1.6 Using BioFrameNet

The specific goal of this project was to create a lexical resource that is relevant to molecular biology. However, linguistic information provided by lexical resources like FrameNet and BioFrameNet can be used in a variety of natural language processing tasks. One such task is that of automatic semantic role labeling (ASRL), as performed, for example, by a system presented by Yi and Palmer (2004). To explore this particular use of BioFrameNet, the original unannotated GRIF data used in this study were run through Palmer's ASRL system. It could perhaps be useful to compare the labeling results of this run with the manual labeling done by biologists, as it may provide insight into the performance of an ASRL tool.

The results returned after the ASRL run are in the form of semantic role labels used by PropBank, a lexical resource similar to FrameNet.[15] Following the schema of the SemLink project[16] outlined by Loper et al. (2007), PropBank argument labeling was mapped to labeling of BioFrameNet. The details of this processing, the results obtained, and problems with handling domain-specific texts like GRIFs that arose are discussed briefly in the final chapter.

1.7 Outline of following chapters

The structure of the rest of the dissertation is as follows. Chapter 2 provides an overview of related lexical resources and domain-specific extensions or adaptations that have been provided for them. FrameNet and PropBank are the two resources focused on most, and adaptations discussed cover two separate domains: molecular biology and soccer.

[14] Usage note: the organizational unit 'class' is often referred to by ontologists as 'frame'. In this work, I will only use 'frame' to refer to units of linguistic semantics.

[15] Similarities and differences between FrameNet and PropBank are discussed in detail in chapter 2.

[16] SemLink was a project initiated and coordinated by Martha Palmer, at the University of Colorado in 2005.

Chapter 3 provides a detailed introduction and formal specification of BioFrameNet, and provides sets of annotations that illustrate frame semantic analyses of two of the most frequent predicates and associated grammatical structures observed in the GRIF collection covered in this study. Chapter 4 considers differences in structure between an ontologically-grounded knowledge base that offers analysis of texts in this GRIF collection and that of BioFrameNet. These differences involve higher-level units of analysis, proposed relations between them, and motivations for the relations defined. Finally, chapter 5 offers conclusions and suggests directions for further study, including potential use of BioFrameNet in tools for natural language processing tasks.

Chapter 2

Lexical Resources and Extensions/Adaptations

2.1 Introduction

A number of computerized lexical resources have been created over the last couple decades, including WordNet (Miller 1995; Fellbaum 1998), VerbNet (Kipper 2005), Lexical Conceptual Structure (Dorr 2001), and Explanatory Combinatorial Dictionary (Mel'čuk 1998; Zholkovsky 1984). These resources show a number of important differences in structure, organization, and content. Most importantly for this study is the extent to which syntactic and semantic interaction is analyzed and recorded.

Lexical resources like these can be extended to or adapted for specific domains. For example, WordNet extensions have been proposed for several domains, including architecture (Bentivogli et al. 2004) and medicine (Smith and Fellbaum 2004; Fellbaum, Hahn, and Smith 2006; Buitelaar and Sacaleanu 2002). While the syntactic and semantic complexities analyzed in the base resource are present in domain extensions, special terminology and language of specific domains can introduce additional complexities.

In this chapter I present a detailed description of two specific lexical resources, FrameNet and PropBank. These are singled out for discussion for two reasons: first, they both focus on syntactic and semantic combinatorial possibilities of predicates and their arguments, and second, a domain extension has been created for each of them. These domain extensions will also be described. In both cases, similarities and differences between the base resource and its extension are highlighted.

In the descriptions provided here, the focus will be on illustrating what sorts of lexical information and analysis are included in the resources.

2.2 FrameNet

As introduced in the previous chapter, FrameNet is a lexical resource based on frame semantic analysis of specific Frames and the lexical units (LUs) that evoke them (Ruppenhofer et al. 2006; Fillmore et al. 2003a, b; Baker et al. 2003). In this context an LU is a pairing of form and meaning, where the form is the lexeme or dictionary form (not distinguishing inflectional variants), and the meaning is the sense associated with the lexeme in a given Frame. (Polysemous words are represented as LUs belonging to different Frames.) FrameNet provides annotated sentences which show how words and phrases that occur in grammatical construction with target LUs in those sentences provide information about components of the Frame.

FrameNet currently includes more than 825 Frames and over 10,000 LUs. The parts-of-speech covered in the collection of LUs include thousands of verbs, nouns, and adjectives, and a much smaller number of prepositions, adverbs, numbers, conjunctions, and interjections. The Frames are organized in a network that includes several kinds of Frame-to-Frame relations. These will be described in section 2.2.2.

FrameNet has provided frame semantic annotations of over 135,000 sentences. The data for this are taken mostly from the British National Corpus, though recently the LDC North American Newswire corpora and the American National Corpus have been used as well.[17] FrameNet's annotation efforts fall into two broad classes: 1) lexicographic annotation, in which specific LUs are chosen as target of analysis, and 2) annotation of running text, in which each Frame-evoking lexical unit that sentences provide is annotated. The annotations discussed in this study are of the first class: specific lexical units were chosen as targets, and then texts that use them were collected. A primary goal of this sort of annotation is to determine and illustrate the syntactic and semantic combinatorial possibilities exhibited by the target lexical units.

The next sections go through three topics: important units of analysis for FrameNet (2.2.1), the kinds of Frame-to-Frame relations FrameNet specifies (2.2.2), then the criteria FrameNet uses to group LUs in a Frame (2.2.3).

2.2.1 Units of analysis

The primary units of analysis FrameNet uses are Frames, frame elements, semantic types, lexical units (LUs), and valence patterns. I will provide here a brief description of these units.

Frames

Though an initial description of a FrameNet Frame was given above, here we focus on definitions of Frames. FrameNet provides Frame definitions through Frame reports for each of its Frames. These reports provide four elements that define the Frame. First, a textual description of the scenario represented by the Frame is provided. Second, a list of the Frame's frame elements is provided. Third, specification of all the Frame-to-Frame relations the Frame is part of is listed. Finally, the LUs that evoke the Frame are listed. Frame-to-Frame relations will be discussed in a later section. In the following paragraphs, I describe FrameNet's definition and categorization of frame elements and semantic types.

Frame elements

Frame elements are representations of the participants and props associated with the scenario of the Frame, and are given Frame-specific labels. All of the LUs of a given Frame use the same set of semantic role labels, thus guaranteeing collection level consistency of semantic role labeling.

[17] Web sites for these corpora: http://www.natcorp.ox.ac.uk, http://www.ldc.upenn.edu, http://americannationalcorpus.org.

Frame elements are specified as one of three types: core, peripheral, and extra-thematic; these are indicators of what is called “coreness” status. Core frame elements are elements that are most central and conceptually necessary for the Frame being defined; these elements bring the Frame uniqueness, a point that will be discussed in chapter 4. Peripheral frame elements introduce items not specific to the Frame, and typically provide instead "circumstantial" information (e.g., expressions of time, place, or manner). These could be used in many other Frames. Extra-thematic frame elements are cases where an element of some other Frame is brought in as a larger embedding context for the Frame’s scenario.

As an example, consider the Frame ‘Commerce_buy’, defined as follows:

(1) Definition for Frame ‘Commerce_buy’

> These are words describing a basic commercial transaction involving a buyer and a seller exchanging money and goods, taking the perspective of the buyer.

The core elements of this Frame are Buyer and Goods. These are conceptually necessary for the successful evocation of the scenario of the Frame. The peripheral frame elements are Means, Money, Rate, Seller, Unit, Place, Purpose, Time, Manner, and Duration. And the extra-thematic frame elements are Reason, Recipient, and Purpose_of_goods.

The following example sentence, with the lexical unit ‘buy.v’ from the Commerce_buy Frame, demonstrates these distinctions:

(2) Myeloski had insisted on buying Duncan a pizza at the latest Pizza Hut .

The four frame elements of the Commerce_buy Frame included in this sentence are Buyer (core), Goods (core), Recipient (extra-thematic), and Place (peripheral). The labeled bracketing in (3) specifies which portions of the sentence are these frame elements’ realizations:

(3) [Myeloski $_{\text{Buyer:core}}$] had insisted on BUYING [Duncan $_{\text{Recipient:extra-thematic}}$]
[a pizza $_{\text{Goods:core}}$] [at the latest Pizza Hut $_{\text{Place:peripheral}}$] .

The Buyer and Goods frame elements in this sentence, ‘Myeloski’ and ‘a pizza’, provide the central participants of the Commerce_buy Frame, while ‘at the latest Pizza Hut’ provides location context for the purchase event. The Recipient frame element, ‘Duncan’, provides the realization of an element of the external Frame Giving.[18] Duncan is not a participant in the commercial transaction described in this sentence, but is the intended recipient of an independent later act of Giving. Sentences with extra-thematic elements typically represent a blending of two or more Frames.

[18] FrameNet does not specify which Frame an external frame element is part of.

Frame element realization

FrameNet annotation consists of highlighting the target LU and indicating which portions of the sentence realize particular frame elements, as illustrated in the previous section. Annotation of frame element realizations includes entire constituents relative to the target lexical unit rather than just the head of those constituents. These are not linked to constituents of a particular parse tree for the entire sentence, as the focus is on a specific target.

Two additional kinds of information about frame element realizations are gathered and stored, namely grammatical function and phrase type. These specify grammatical information about frame element realizations in relation to the target LU.

Null instantiation

For an utterance of a lexical unit to be considered felicitous, its core frame elements must either be realized directly, or be possible to infer. FrameNet annotation identifies both of these cases. In the example sentence given above, two core frame elements were realized directly. For cases of inferred core frame elements, FrameNet analyzes them as a form of "null instantiation".

There are three types of null instantiation: constructional null instantiation, indefinite null instantiation, and definite null instantiation. Constructional null instantiation involves cases of grammatically licensed omission. Common cases for verbal lexical units are passives and imperatives:

(4) The wine was purchased yesterday. Buyer omitted,
licensed by passive construction

Buy us some treats, the kids pleaded. Buyer omitted,
licensed by imperative construction

Indefinite null instantiation involves cases in which a core frame element is given an indefinite or default interpretation. For example, the LU 'eat.v' in the Ingestion Frame allows an indefinite interpretation of the Ingestibles frame element:

(5) We ate at the park. Ingestibles omitted via INI,
given an indefinite interpretation

vs.

We ate sandwiches at the park. Ingestibles directly realized
('sandwiches')

Definite null instantiation involves a kind of zero anaphora: a core frame element is not expressed within the target LU's scope, but must be inferable from either the linguistic context or background knowledge. In the following examples, the portion in parentheses specifies conditions that would license the omission of a core frame element:

(6) This one is similar (to something known in the conversation).

```
Similarity: Entity_2
```

Let me explain (a mystery that we've been talking about).

```
Explaining_the_facts: Question
```

When did she arrive (at the place we all have in mind)?

```
Arriving: Goal
```

Who won (the contest that's been a part of our shared conversation or context)?

```
Finish_competition: Competition
```

An important distinction between these versions of null instantiation is that while constructional null instantiation is licensed by constructions of the grammar and thus available to any LU that can participate in such constructions, indefinite and definite null instantiation are licensed by particular lexical units. This lexical specificity creates different grammatical possibilities for LUs even within the same Frame. For example, though 'eat.v' and 'devour.v' both belong to the Ingestion Frame, only 'eat.v' licenses the indefinite null instantiation shown above:

(7) We ate at the park.
*We devoured at the park.

FrameNet notes instances of null instantiation in annotations, specifying which frame element is omitted and by which sort of null instantiation.

Semantic types

To indicate selectional requirements, frame elements can be assigned a semantic type. Occasionally LUs that are opposites belong to the same Frame. Semantic types allow for such distinctions to be noted. For example, 'praise.v' and 'criticize.v' both belong to the Judgment_communication Frame. Whether the judgment implied is positive or negative can be specified in the Semantic Type given to the Evaluee frame element of this Frame, either Positive_judgment or Negative_judgment.

Another important use of Semantic types is the categorization of and constraints on the fillers of a frame element, i.e., its realization. For example, frequently a Frame's scenario requires that a frame element must be sentient. This requirement is taken care of with a semantic type specification.

Semantic types were originally developed through careful semantic analysis by the FrameNet team. In order to attain consistency in such typing, efforts have been made recently to link semantic types with ontological resources (Dolbey et al. 2006). As will be discussed in the next chapter, BioFrameNet also does just this sort of linking with domain ontologies.

Lexical Units
These are words and multiword expressions, each of which belongs to a particular Frame. For each lexical unit, FrameNet provides two separate reports. In one of them, the Annotation Report, all of the annotated sentences which demonstrate the lexical unit are displayed. The target lexical unit is marked and the portions of the sentence that are the realization of a frame element are highlighted in the particular color designated for that frame element.

In the other report, the Lexical Entry Report, an overview of the grammatical properties of frame elements' realizations is provided. An overview like this is a helpful way of summarizing the syntactic and semantic combinatorial possibilities of the predicates that evoke a Frame, one of FrameNet's key goals. This kind of report is an important part of the domain extension described in this study, and will be discussed further in the next chapter.

Valence patterns
The final unit of analysis for FrameNet covered here is **valence pattern**, a specification of syntactic and semantic combinations used with LUs for frame element realizations or omissions. Given a sentence that includes a lexical unit of a particular Frame, valence patterns specify the grammatical structure of the frame elements realized in this sentence, along with information about the omissibility of any core elements via some sort of lexically determined null instantiation. Because frame element omissibility varies across LUs even within the same Frame, valence patterns do not generalize across Frames, but rather are particular to LUs.

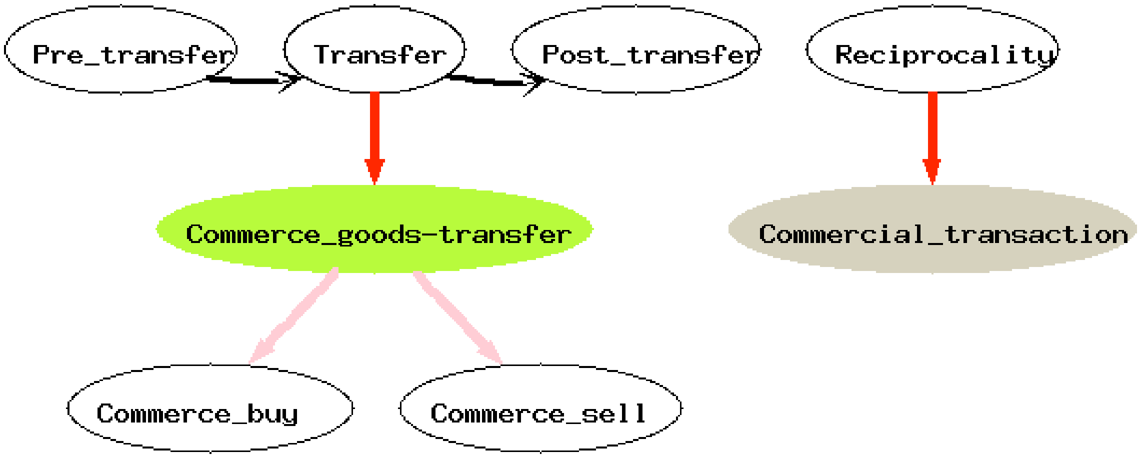
By collecting different valence patterns used by particular LUs, useful summaries of the syntactic and semantic combinatorial possibilities for them can be produced, an important goal for FrameNet.

2.2.2 Frame-to-Frame relations

FrameNet defines several ways of relating Frames to other Frames. I will describe two specific types of relations here.

One type of Frame-to-Frame relation is **Inheritance**, in which one Frame is an elaboration of another Frame. An example of this is Commercial_transaction Frame. This Frame inherits from Reciprocality, a background Frame that emphasizes relations between two separate parties. The elaborations in this case are the relations of trading money for goods and the agreed change of ownership that is a result of the trade. Another inheritance example is the Transfer Frame being inherited by the Commerce_goods-transfer Frame and in doing so, characterizes more fully the before and after states of a transfer event. In the following diagram of these Frames, the red arrows illustrate instances of the Inheritance relation:

(8) Frame-to-Frame Relations



Another type of Frame-to-Frame relation is **Perspective_on**, in which one Frame imposes a point of view on another Frame. Examples of this relation are the Commerce_buy and Commerce_sell Frames, each of which imposes a different perspective on the Commerce_goods-transfer Frame. The perspective Commerce_buy imposes focus on the buyer, while the perspective Commerce_sell imposes focus on the seller. A similar relation exists between the Commerce_money-transfer Frame and the different perspectives for it imposed by Commerce_pay and Commerce_collect. The pink arrows in the previous diagram show instances of the Perspective_on relation for Commerce_goods-transfer with Commerce_buy and Commerce_sell. Similar instances of this relation exist between Commerce_money-transfer and its two perspectivizing Frames, Commerce_pay and Commerce_collect.

There are several other types of Frame-to-Frame relations defined in FrameNet. The most important one for this study is **Causative_of**, a relation that will be discussed in the next chapter.

2.2.3 Criteria for grouping

In deciding which lexical units belong together in the same Frame, FrameNet considers a variety of criteria based on the semantics of the Frame, especially the scenario evoked, the participants and props involved, and perspectives of the participants. These criteria will be discussed further in chapter 4, which focuses on decisions made about creating the BioFN Frames that targeted bio-predicates belong in.

2.3 PropBank

PropBank (Palmer et al. 2005; Kingsbury and Palmer 2002) is a lexical resource whose original goal was to add a semantic layer of predicate argument structure to the

TreeBank's syntactic parses of several Wall Street Journal volumes, a corpus of one million words (Marcus 1994). It is well known that for many important NLP tasks (e.g., information extraction, question answering, translation), a syntactic parse alone does not provide enough information for successful completion of these tasks. Two of the most common problems are that there can be identical syntactic structure for clauses with different meanings, as seen in the following examples:

(9) The sergeant **played** *taps*. direct object is piece performed
The sergeant **played** *a beat-up old bugle*. direct object is instrument used

and the same meaning can be provided by several different syntactic structures:

(10) John will **meet** with Mary.
John and Mary will **meet**.

Cases like these demonstrate that NLP tools need information about the semantics of the data being analyzed for successful processing of these data. Predicate argument structure serves as a useful starting point for this.

An important goal in creating PropBank was to annotate enough cases of variation in predicate argument realization to allow for successful machine learning of semantic role labeling. In showing realizations of semantic roles in different syntactic alternations, this resource also offers an opportunity to measure the frequency of particular alternations seen in real-world, domain-independent textual data. FrameNet, by contrast, makes no attempt to have annotations that reflected frequency differences.

In the following sections, I will describe the formal representation of predicate argument structure PropBank defines (2.3.1), and some general differences with FrameNet (2.3.2).

2.3.1 Formal representation of predicate argument structure

Three important units of analysis for PropBank are rolesets, syntactic frames, and framesets. In this section I will describe these items.

Rolesets

For each verb usage, PropBank defines a **roleset**, a set of verb-specific semantic roles for the verb's arguments.[19] These are given numbered labels, starting with zero: arg0, arg1, arg2, up to arg5. They are also given corresponding mnemonic labels. For example, this is the roleset for the verb 'accept.01', "take willingly":

[19] The notion "verb usage" is a correlate of FrameNet's lexical unit.

(11) PropBank roleset for ‘accept.01’

Arg0: acceptor
Arg1: thing-accepted
Arg2: accepted-from
Arg3: attribute

(12) [Mary $_{\text{arg0}}$] **accepted** [a gift $_{\text{arg1}}$] [from John $_{\text{arg2}}$].

In addition to semantic roles defined in the rolesets, PropBank also annotates adjuncts of a variety of sorts (e.g., location, purpose, manner). These are very much like frame elements categorized as peripheral in FrameNet.

For cross-verb generalizations, PropBank attempts to make consistent choices of role labels based on event relations across verbs’ arguments. For example, Arg0 is chosen for the argument that is the agent. If a verb doesn’t have an argument with an agent-like semantic role, Arg0 is not used. For example, the roleset for the verb ‘fall.’, “move downward”:

(13) PropBank roleset for ‘fall.01’

Arg1: logical subject, patient, thing falling
Arg2: extent, amount fallen
Arg3: start point
Arg4: end point, end state of arg1

(14) [Profits $_{\text{arg1}}$] **fell** [by 3% $_{\text{arg2}}$].

VerbNet’s grouping of verbs in classes is a useful guide for establishing and verifying consistency in other cases of role assignment in PropBank.

Syntactic frames

Each verb is associated with a set of **syntactic frames**, i.e., patterns of argument realization. For example, considering ‘accept.01’, a verb example discussed in the previous section, the following are alternative syntactic frames available for this verb:

(15) PropBank roleset for ‘accept.01’, “take willingly”

Arg0: acceptor
Arg1: thing-accepted
Arg2: accepted-from
Arg3: attribute

(16) [Mary $_{\text{arg0}}$] warmly **accepted**.
[Mary $_{\text{arg0}}$] **accepted** [a gift $_{\text{arg1}}$].
[Mary $_{\text{arg0}}$] **accepted** [a gift $_{\text{arg1}}$] [from John $_{\text{arg2}}$].
[Some practices $_{\text{arg1}}$] are **accepted** [as normal $_{\text{arg3}}$] here.

Cataloging the various syntactic frames associated with different verbs is one of the main information structures VerbNet provides.

Frameset

The final unit of analysis considered here is **frameset**, a grouping together of the roleset and associated syntactic frames for a given verb sense. PropBank offers one of these for each distinct verb sense.

Though in general each distinctive verb sense gets its own frameset, an important analytical choice for PropBank is determining when to create separate framesets for a different uses of the same lexeme. This is similar to the choices FrameNet makes for deciding whether or not separate lexical units belong in the same Frame, only in this case the choice is made at the scope of a single lexeme. Differences between FrameNet's and PropBank's treatment of this issue will be discussed in a later section (2.3.2).

Like FrameNet, PropBank uses both semantic and syntactic criteria for making this decision. The most important criterion is whether or not the uses take a different number of arguments. For example, consider the lexeme 'decline'. Two uses of this lexeme have undeniably distinct meanings, and the different number and kinds of semantic roles realized when they are used corresponds to this difference in meaning:

(17) PropBank rolesets for two different senses of lexeme 'decline'

a. 'decline.01', "go down incrementally"

Arg1: entity going down
Arg2: amount gone down by
Arg3: start point
Arg4: end point

[Net income $_{\text{arg1}}$] **declined** [0.2% $_{\text{arg2}}$] [to \$10 million $_{\text{arg4}}$].
vs.
b. 'decline.02', "turn down"

Arg0: entity turning down
Arg1: thing turned down

[John $_{\text{arg0}}$] **declined** [comment $_{\text{arg1}}$].

Because of differences in the number of roles and the semantics of the roles, these two uses are given separate framesets.

Another kind of case where PropBank posits separate framesets is verb-particle constructions like the following:

(18) PropBank rolesets for verb particle cases 'cut', 'cut off', 'cut back'

a. 'cut.01', "slice"

Arg0: cutter
Arg1: thing cut
Arg2: medium, source
Arg3: instrument
Arg4: beneficiary

b. 'cut.04', "cut off = slice"

Arg0: cutter
Arg1: thing cut
Arg2: medium, source
Arg3: instrument
Arg4: beneficiary

c. 'cut.05', "cut back = reduce"

Arg0: cutter
Arg1: thing reduced
Arg2: amount reduced by
Arg3: start point
Arg4: end point

Each of these is given its own frameset, even in the cases of cut.01 and cut.04 which otherwise exhibit complete similarity in the semantics of their rolesets.

In contrast to the cases just illustrated, certain other cases of differences in rolesets are argued to be alternations that preserve meanings and thus not warrant creating a separate frameset. Typical cases here are causative and inchoative alternations such as the following different uses of 'open':

(19) PropBank roleset for two different senses of lexeme ‘open’

‘open.01’, “open”

Arg0: opener
Arg1: thing opening
Arg2: instrument
Arg3: benefactive

[John $_{\text{arg0}}$] **opened** [the door $_{\text{arg1}}$].
[John $_{\text{arg0}}$] **opened** [the door $_{\text{arg1}}$] [with his foot $_{\text{arg2}}$].
[John $_{\text{arg0}}$] **opened** [the door $_{\text{arg1}}$] [for Mary $_{\text{arg3}}$].
[The door $_{\text{arg1}}$] **opened**.

For PropBank, the lack of particular arguments such as agent, instrument, or benefactive in these different uses is argued to be cases of arguments being left unspecified, and thus separate framesets are not posited. This is similar to FrameNet’s handling of variation in lack or presence of peripheral frame elements, in that these are not determinative factors of Frame membership. By contrast, however, the beneficiary argument "for Mary" is regarded in FrameNet as extra-thematic, since the intention to do this on someone's behalf is not a part of the core meaning of "open".

Another case when different framesets are not proposed is when there are differences in the syntactic type of realizations of semantic roles:

(20) Differences in syntactic type of realization of semantic role, for lexeme ‘decline’

‘decline.02’, “turn down”

Arg0: entity turning down
Arg1: thing turned down

[John $_{\text{arg0}}$] **declined** [comment $_{\text{arg1}}$]. arg1 is NP
[John $_{\text{arg0}}$] **declined** [to elaborate $_{\text{arg1}}$]. arg1 is infinitival VP

Both uses here are covered in the same frameset. FrameNet is similar here in that grammatical characteristics of frame element realizations is not a relevant consideration when determining whether or not a particular lexical unit belongs in a different Frame, or perhaps belongs in two separate Frames as in cases of polysemy.

Frames file

For each verbal lexeme in the Wall Street Journal corpus, PropBank places all of the lexeme’s framesets, along with examples of their realization, in a dedicated file called a **Frames file**.[20] There are over 3300 verbs included in this corpus, and following the

[20] Potential point of confusion: the terms “frameset” and “frames file” in PropBank are unrelated to frame semantics or FrameNet.

criteria discussed above for determining which verb uses require a new frameset definition, over 4500 framesets were created and stored in the separate frames files. These were used for annotation of predicate argument structures observed in the corpus.

2.3.2 Differences with FrameNet

In this section I will describe key differences between PropBank and FrameNet. Though the two resources share the goal of semantic annotation, there are some important differences in their approach, including 1) data selection, 2) linking of syntax and semantics, and 3) primary unit of analysis. These are discussed below.

Data selection

The most noticeable difference between the two is that PropBank provides semantic analysis only of verbal lexical items, while FrameNet includes semantic analysis of all major parts of speech. This offers FrameNet on one level greater breadth of lexical coverage.

On the other hand, another important difference between the two is selection of verbal predicates analyzed, and selection of sentences for annotation. FrameNet selects specific verbal lexical units based on Frames being created, and chooses a small subset of sentences to annotate for them, ones that demonstrate as great a variety possible in syntactic and semantic combinations. By contrast, PropBank annotates all of the clauses in the corpus they choose, regardless of verb sense or complexity of structure that the verb is used in. This offers PropBank greater breadth of coverage for verbal lexical items, and results then in coverage of structural complexity not necessarily encountered in FrameNet's annotation collection.

Linking syntax and semantics

PropBank annotations of predicate argument structure make reference to specific tree nodes of TreeBank's parses of the corpus data. By contrast, as described earlier in section 2.2.1, FrameNet annotations are not linked to syntactic parse trees. Instead, markers in annotations are based on character offsets of frame element realizations in source texts, and often cover two or more constituents. The text included then often does not correspond to syntactic constituents provided by a syntactic parse of the whole sentence. This difference could create difficulties for end users who want to perform automatic processing that includes information from FrameNet's annotation collection.

Primary unit of analysis

One final distinction to note is the different primary unit of analysis chosen in these two resources. For PropBank, this is individual verbal lexical items and their framesets. For FrameNet, it is Frames and the collection of lexical units included in them. There are several implications associated with this difference.

First, as noted in section 2.2.1, collection level consistency of semantic role assignment is obtained automatically for FrameNet by the definition of frame elements in Frames, thus assuring that separate lexical units within the same Frame will have consistent uses of semantic roles. Further, the Frame-by-Frame relation of inheritance includes a mapping of frame elements across Frames and thus further guarantees consistency in role assignment. For this sort of collection level semantic analysis, PropBank must use external means such as considering VerbNet's definitions of verb classes.

Second, PropBank and FrameNet follow different criteria for determining when a new instance of the primary unit of analysis is warranted. For PropBank, creation of a new frameset is largely a question of whether or not there are differences in semantic roles used. For FrameNet, creation of a new Frame is determined based on a set of criteria, mostly having to do with the semantics of Frame definitions. Differences arise here in a few cases mentioned in section 2.3.1, namely PropBank's treatment of causative/inchoative alternations and of verb-particle constructions.

In cases of causative/inchoative alternations, PropBank does not create a new frameset, but instead assumes a frameset that holds all possible roles (e.g., agent, instrument, etc.), and considers omissions of any of these arguments to be due to their being left unspecified. By contrast, if any systematically omitted arguments are categorized as core frame elements, then a new Frame must be created in which the omitted elements are not included in the list of core elements.

For verb-particle constructions, PropBank always creates a new frameset for them, regardless of whether or not they have similar semantics. By contrast, whether or not a new Frame is considered necessary for them completely depends on the semantics they evoke.

For further analysis of these differences in primary unit of analysis between PropBank and FrameNet, see Ellsworth et al. (2004).

2.4 Domain-specific FrameNet extension: Kicktionary

Kicktionary is a multilingual extension of FrameNet for the specialist domain of soccer, created by Thomas Schmidt (presented in Schmidt 2008).[21] It applies most of the FrameNet techniques and terminology for identifying participants, roles, events, results of events, and the forces that foster or prevent such events, and special terminology for

[21] Schmidt uses the term 'football', not 'soccer'.

them. Events of soccer are different from most events covered in FrameNet in that they are "institutional", i.e., events regulated by stipulated objectives and constraints, the rules of the game, rather than determined empirically by observation.[22] Parallel lexical analysis for soccer language in English, French, and German is provided. Schmidt's primary goal for creating Kicktionary was to produce a resource for assisting *human* users in understanding, translating, and paraphrasing language of soccer.

Schmidt proposes a set of scenes, Frames, and lexical units specific to this domain.[23] Scenes are conceptual entities about the events of soccer. Frames, on the other hand, provide frame semantic entities that host mapping of scenes and their linguistic realization via lexical units and frame elements. Each Frame is specified as belonging to a particular scene, and each lexical unit is specified as belonging to a particular Frame. As with FrameNet and BioFrameNet, frame elements are defined for each Frame. For example, here is one scene and two Frames that belong to it, and two lexical units that belong to the Frame:

Scene: Pass. The Pass scenario is centered around the event of a player transferring the ball to a team-mate. The main protagonists of the scenario are the passer and the recipient. Using a part of his body, the passer directs the ball towards the recipient. The ball moves in a certain direction from the source location on the field along a path to a target location thereby covering a certain distance.

(21) Kicktionary: FrameNet Extension for specialist domain of soccer

Frame:	Control
Frame Elements:	Recipient, Pass, Ball, Target, Part_of_body, Passer
Lexical units:	trap.v, miss.v
Frame:	Intercept
Frame Elements:	Interceptor, Pass, Intervention_location, Ball, Target
Lexical units:	interception.n, misjudge.v

Belonging to the same scene is the only Frame-to-Frame relation kicktionary offers. Therefore there are no Frame-to-Frame relations between Frames of kicktionary and those of FrameNet.

As with FrameNet and BioFrameNet, analysis of the Frames and lexical units of this domain is demonstrated by annotations of textual corpus data, in this case publicly available soccer match reports. For this resource, corpus data in each of the three covered languages is used. Kicktionary also includes annotation of spoken match commentary that has been transcribed. These annotations indicate where in the sentences each of the frame elements is realized (if realized at all). Unlike FrameNet, further grammatical information about frame element realizations was not collected, nor was information about cases of null instantiation.[24] However, tables with useful overview and

[22] The closest to this in FrameNet would be the Frames Arraignment and Criminal_process.

[23] Schmidt proposes a separation of conceptual scenes and linguistic Frames.

[24] These were not included due to lack of time, not because they were regarded as unimportant.

summary information about frame element realizations for each lexical unit in each Frame are provided. Here are example annotations for two of the lexical units listed above.

(22) Kicktionary: example annotations
Frame: Control, LU: *trap.v*

On the half-hour [Edgaras Cesnauskis $_{\text{Recipient}}$] **trapped** [a long, lofted pass $_{\text{Pass}}$] and rounded the exposed Iker Casillas, only for Carles Puyol to clear off the line.

Frame: Intercept, LU: *interception.n*

After just three minutes, [veteran striker Gert Verheyen $_{\text{Interceptor}}$] *made* a fine **interception** [in midfield $_{\text{Intervention\_location}}$] and passed to team-mate Rune Lange.

Schmidt notes that more than half the lexical units in kicktionary are nominal. In many of these cases, frame elements are realized externally via support verbs. As with FrameNet and BioFrameNet, annotation for these cases includes marking of the support verb. This was shown in the example annotation for **interception**, in which the support verb ‘make' is italicized. Another result of the frequent use of nominal lexical units is that frame element realization often involves complex multiword expressions with combinations of compounds and possessives. Unfortunately, because of lack of grammatical analysis of them, we do not get an overall picture of how complex expressions like these realize frame elements and their combinations.

In addition to spoken language transcriptions and multiple languages covered, kicktionary includes other items that are not part of FrameNet or BioFrameNet. In particular there are WordNet-inspired (Miller 1995; Fellbaum 1998) synsets and accompanying hierarchies of semantic relations like synonymy and troponymy. Also, kicktionary provides visual representations of scenes with diagrams and pictures.

There are several things that are part of FrameNet and BioFrameNet but are not included in kicktionary. There are no coreness distinctions in kicktionary. And as was mentioned above, detailed grammatical information about frame element realizations is not included. Another item not provided is specification of null instantiation of frame elements. One consequence of these omissions is that there are no lexical entry reports for soccer predicates available. Yet it is these reports that assist in achieving an important goal for FrameNet and BioFrameNet, namely exploring and analyzing syntactic and semantic combinatorial possibilities for evoking semantic frames.

2.5 Domain-specific PropBank extension: PASBio

PASBio is an extension of PropBank for the specialist domain of molecular biology (Wattarujeekrit et al. 2004; Wattarujeekrit and Collier 2005; Wattarujeekrit 2005; Cohen and Hunter 2006)[25]. It adds predicate argument structure framesets for verbs used to describe and characterize domain-specific events.[26] Verbs that cover a few specific kinds of biological events were chosen as starting points.[27] The primary goal of this resource extension is the same as that for the base it is extending: create a resource that is useful for building successful NLP tools, especially ones for automatically extracting information about biological events.

Events of this domain involve molecular entities that participate in the completion of a process. As Wattarujeekrit et al. (2004) point out, these events often involve some sort of transformation of one or more of these entities or of the cellular region where they are located. Consequently the arguments used with the verbs chosen for analysis are typically references to domain specific entities, in particular genes and gene products. Frequently there is also mention in the same sentence of overall effects of the events the verbs describe.

For automatic processing of the language used in this domain, there are problems similar to the ones discussed earlier for processing of language used in other domains, and for that matter, language not specific to any particular domain. Namely, the same syntactic structure can mean different things, and the same meaning can be expressed in different ways. And yet it is sometimes argued that the complexity of biological terminology and that of the sentences structure used to describe events in this domain provide an even greater challenge for automating information extraction (Wattarujeekrit et al. 2004; Wattarujeekrit 2005; Friedman 2002, referencing Harris 1982, 1991; but see also Wermter and Hahn 2004).

The texts used in building framesets for PASBio were MEDLINE article abstracts and also complete journal articles. There was a much greater variety of text sources used in building PASBio (the different journals that contained articles and abstracts) than there was for PropBank. However, a similar level of diversity might have been obtained for PropBank because the Wall Street Journal corpus it used included articles written by many different authors.

Examples of framesets defined in the PropBank served as models for defining PASBio framesets in the extension. Verbs were chosen based on their frequency in the texts collected, and on their importance in the initial targeted event types. For each verb, sample sentences with the verb's usages were collected to define its frameset. Sentences were chosen so as to cover as broad a range possible of usages of each verb, and in this sense, the same strategy was followed as was for the resource base, PropBank. One important difference, though, is that manual creation of domain framesets required domain expert introspection and evaluation.

In addition to the kinds of language analyzed in building framesets, another important difference between PASBio and PropBank involves definition of rolesets. PASBio

[25] The following is the URL for a PASBio website: http://research.nii.ac.jp/~collier/projects/PASBio/

[26] Wattarujeekrit et al. call these ‘frames’, despite PropBank’s different use of this term.

[27] The first event types targeted are gene expression, molecular interactions, signal transduction

defines two different kinds of core arguments. One kind is for arguments that play a role *during* the event described by the verb. The other kind is for arguments that play a role *after* the event, and typically express results or consequences of the event, information that is useful for understanding the importance of the event described. The second kind is similar to arguments included as purpose adjuncts in PropBank, except that for PASBio they are considered core. Here we see a case where domain semantics and the goal of information extraction drive the inclusion of items separate from the primary event described by the predicate itself. For FrameNet, this sort of addition to the semantic role set would most likely be analyzed as an extra-thematic frame element, as it is semantically outside of the scope of the particular Frame evoked by the predicate.

For notation, the first kind of core argument is labeled as it is in PropBank, with 'Arg' plus a number (Arg0, Arg1, etc.). As with PropBank, Arg0 is reserved for agent like roles. The other kind of argument is labeled with ArgR. The following example from PASBio's website[28], using the verb 'mutate.01', shows annotations of these different kinds of roles:

(23) PASBio roleset for 'mutate.01', (WordNet Sense 1 "undergo mutation")
Arg1: physical location where mutation happens // exon, intron //
Arg2: mutated entity // gene //
Arg3: changes at molecular level /* always use prepositional phrase */
ArgR: changes at phenotype level /* secondary predication */

The [exon 5 $_{\text{arg1}}$] **mutated** [allele $_{\text{arg2}}$] [with the premature translation termination $_{\text{arg3}}$] [resulted in severe deficiency of Hex A $_{\text{argR}}$].

In this example there are three instances of the first kind of core argument used and a single instance of the second kind.[29] For PropBank, the sort of structure shown in this particular ArgR realization would most likely not be included as a semantic role for the preceding verb participle.[30] FrameNet wouldn't even include this element as an instance of an extra-thematic frame element, as it is not part of the local grammatical structure of the predicate 'mutate'.

Both of the core argument types shown above are separate from adjunct like items which are covered the same way for PASBio as they are for PropBank, as the following example shows:

[28] PASBio site's URL: http://research.nii.ac.jp/~collier/projects/PASBio/

[29] The verb form in this sentence is the participial version of the verb being analyzed, and is a portion of a complex compound noun phrase, a typical example of complex compounds used in this domain.

[30] PropBank doesn't yet include annotations of verb participles.

(24) PASBio roleset for ‘abolish.01’, (WordNet Sense 1 "do away with")
Arg0: causer/agent // change at molecular level //
Arg1: entity abolished // normal transcription, normal splicing //
ArgM-ADV: adjunct[31]

[Mutation of the proximal GATA element in the context of the –700 or the –135 bp promoter $_{\text{arg0}}$] [completely $_{\text{argM-adv}}$] **abolished** [synergy $_{\text{arg1}}$], indicating that this element is essential for MEF2 –GATA-4 cooperation.

PASBio pays special attention to adjuncts in which a biologically important aspect of the event is asserted, as in the previous example’s use of ‘completely’.

2.6 Summary

This chapter has presented a description of two specific lexical resources, FrameNet and PropBank, and of a domain specific extension for each of them. The descriptions covered the primary units of analysis of these resources, with a focus on the ways in which syntactic and semantic combinatorial possibilities of predicates and arguments are represented and analyzed.

The next chapter will introduce BioFrameNet, an extension of FrameNet to the domain of molecular biology.

[31] Though ArgM adjuncts are mentioned in Wattarujeekrit et al. (2004), the PASBio website does not state explicitly what sort of adjunct ArgM is in this example. The general English use of ‘completely’ as a degree modifier is likely the appropriate interpretation for its use in this example.

Chapter 3

BioFrameNet

3.1 Introduction

In this chapter I present BioFrameNet (BioFN), an extension of FrameNet (FN) to the domain of molecular biology. The chapter starts with a definition of the structure of BioFN, including how Frames are defined and how texts are annotated using these Frame definitions. Then, a case study demonstrating a use of BioFN is shown. The Frames presented in the case study are two that have been defined for the event of intracellular protein transport. Texts that discuss this sort of event have been annotated using the Frames. A lexicographic sample of several of the Lexical Units (LUs) in these Frames shows some of these annotations.

3.2 Resource structure

The core structure of BioFN is the same as that of FN, described in the previous chapter. For individual Frames, a description of the scenario evoked by the Frame is provided, along with a list of the Frame's frame elements (FEs) and their definitions, and a specification of any relations the Frame holds with other existing Frames. In addition, a list of LUs that evoke the Frame is provided. A Frame Report provides all of these items. With these in place, texts that contain one of the LUs can then be given annotations using definitions of the Frame.

The annotations produced in BioFN follow FN's guidelines for lexicographic annotation, described in Ruppenhofer et al. (2006). Most importantly, annotations are directed toward a single target LU and its dependents. For each sentence annotated, BioFN marks the target LU, and collects and records syntactic and semantic information about the relevant Frame's FEs. The full phrase of each FE realization is marked, not just the head of the phrase. For each FE, three kinds of information are gathered. The first kind is the identity of the specific FE, something that is recorded in all cases. But beyond this, the other kinds of information gathered depend on whether the FE is linguistically realized, or rather, is omitted via some sort of null instantiation. In cases when the FE is explicitly realized, the phrase type (PT) and grammatical function (GF) of the realization are recorded. In cases when the FE is omitted, the type of its null instantiation is recorded. Together these items describe the syntactic and semantic combinations observed in particular instances of FE realizations in a given sentence, in addition to the kinds of FEs that are left implicit. As with FN, this is collected so that the syntactic and semantic combinatorial possibilities of a LU and its associated FEs can be analyzed. Results of this analysis are used in specifying valence patterns for the LU.

Most of the constituents annotated as FE realizations, with the exception of subjects, are constituents that are part of the maximal phrase headed by the target LU. There are, however, two kinds of cases in which non-local constituents are also annotated. These are described in detail in Ruppenhofer (2006, pp. 19-61).

The first kind involves cases in which the target LU is syntactically governed by a support predicate. If a FE is also realized as a dependent of the support predicate governing the target LU, the valence properties of the support predicate guarantee that the argument that realizes the FE is also interpreted as an argument of the target LU. In this context, the non-local constituent is annotated the same way as the locally realized constituents are. The following sentence illustrates an example of this:

(1) recoverin *undergoes* intracellular translocation

In (1), the entity 'recoverin' is an argument of the support verb 'undergo' and is also interpreted as a dependent of the noun 'translocation', and would be given a standard FE annotation. Other support predicates used in the texts analyzed in this study include 'require', 'induce', and 'act'. These predicates also guarantee interpretations of non-local arguments for target LUs.

The second kind involves cases in which the target LU is embedded in a relative clause. In these cases, both the antecedent and the constituent containing the relativizer are annotated as FEs. Figure (2) show an example of this:

(2) the COOH-terminal fragment of ErbB-4 that translocates to the nucleus

In (2), the target LU 'translocate' is embedded in a relative clause that modifies the constituent 'the COOH-terminal fragment of ErbB-4'. This constituent and the relativizer are both given FE annotations.

Constituents that are understood only through context as referring to the filler of a particular frame element role are analyzed as cases of null instantiation (NI), an analysis that is indicated by italicizing the words that could be interpreted as a FE filler, and displaying it in the color assigned to the FE role. Figure (3) shows an example of a text that would be annotated this way:

(3) Analysis of murine CD28 mutants reveals a correlation between TRANSLOCATION to lipid rafts and costimulation of IL-2 production .

In (3), the LU 'translocation' is not part of a grammatical structure in which an accurate interpretation of the identity of the entity involved in translocation is guaranteed. This particular text is offered as a description of the function of the protein 'IL-2'. However, the protein that is involved in translocation is 'murine CD28 mutants'. The grammatical structure headed by the LU does not guarantee an interpretation of either of these proteins as the entity involved in translocation. In order to conclude accurately which analysis is accurate, background knowledge or surrounding context would be necessary. Cases like this are thus given an analysis of NI (Ruppenhofer et. al. 2006, p. 21).

FE annotation

When recording grammatical information for cases when FEs are explicitly realized in a sentence, the PTs and GFs used in BioFN are taken from FN. Table 3.1 below provides abbreviations and names of these items.

Phrase Types	
N	Bare Noun
A	Bare Adjective
NP	Noun Phrase
PP	Prepositional Phrase
Poss	Possessive

Grammatical Functions	
Ext	External Argument
Obj	Object
Dep	Dependent
Gen	Genitive Determiner

TABLE 3.1: Phrase Types and Grammatical Functions in BioFN

The most frequent PTs observed in the data considered in this study are the two phrasal categories *noun phrase* (NP) and *prepositional phrase* (PP). The following sentence, with the target noun LU ‘translocation’, shows examples of these[32]:

(4) Phrase types of phrasal categories: NP, PP
results suggest that [GLUT4 $_{\text{NP}}$] requires TRANSLOCATION [to the plasma membrane $_{\text{PP}}$]

Cases in which the constituent is not a full phrase include non-maximal noun (N) or adjective (A), or possessive (Poss). FEs annotated as N or A are typically either compound pre-modifiers of a target LU or modifiers of a full NP. FEs annotated with PT Poss are either possessive pronouns or noun phrases marked with *'s*. The following sentences show examples of these:

[32] Examples like these are taken from portions of the GRIFs considered in this study. Complete versions of GRIFs will be provided in a lexicographic sample in section 3.3.4.

(5) a. Phrase type of bare noun: N
an electrostatic switch that controls the [membrane $_N$] TRANSLOCATION of the protein

b. Phrase type of bare adjective: A
[juxtanuclear $_A$] TRANSLOCATION of protein kinase C betaII is selectively inhibited

c. Phrase type of possessive: Poss
chemical anoxia activates c-Src and induces [its $_{Poss}$] TRANSLOCATION to cell-cell junctions

The remaining type of information recorded for constituents that explicitly realize FEs is their GF. For BioFN and FN, GF specifications are descriptions of the grammatical role of the constituent with regard to a target LU, not for the full sentence[33]. The four GFs observed here are external (Ext), object (Obj), dependent (Dep), and genitive (Gen). Two of these, Ext and Dep, deserve special mention. Ext is the GF assigned to any item that falls outside the maximal phrase headed by the target LU, most frequently the subject of a target verb, a constituent that controls the subject, or the subject of a support verb that governs a target noun. The following sentences, again with target noun LU 'translocation', or with target verb LU 'translocate', show examples of these:

(6) Grammatical function of external: Ext
[Synembryn $_{Ext}$] TRANSLOCATES to the plasma membrane
[Nm23-H2 $_{Ext}$] was induced to TRANSLOCATE to the plasma membrane
[Sam68 $_{Ext}$] undergoes activity-responsive TRANSLOCATION to the soma

As discussed early in this section, the choice of whether to analyze a non-local constituent as a FE realization with GF Ext, or rather, as an instance of null instantiation, is based on whether or not it occurs in a grammatical structure which guarantees an accurate interpretation of the constituent.

Dep is the GF assigned to complements with PTs of PP, N, or A. In cases in which GF is N or A, the constituent is typically a pre-modifier in a compound with the LU. Examples of FEs with GF Dep are shown below:

(7) Grammatical function of dependent: Dep
SNARE proteins are involved in late steps of [GLUT4 $_{Dep}$] TRANSLOCATION
[membrane $_{Dep}$] TRANSLOCATION [of Dvl-1 protein $_{Dep}$]

As with FN, specification of GF Dep does not imply obligatoriness or optionality of the FE.

When a FE is omitted from the sentence, the instance of null instantiation (NI) is recorded. Cases of NI are very common in the collection of Gene References into

[33] If the target LU is the main verb of the sentence, then the GF specification would be the same as that for the full sentence.

Function (GRIF) texts considered in this study, described briefly in chapter 1. The exact same sort of NI is also seen in abstracts for journal articles. This is unsurprising for two reasons. First, frequently the source of text for a GRIF is the abstract of the journal article which the GRIF is associated with. Second, maximum text length is a common artificial constraint for both GRIFs and journal article abstracts. The two most visible effects of this constraint are the use of complex nominal and verbal compound structures and omission of FE realizations.

A distinction is made in FN between omissions that are grammatically licensed and those that are lexically specific. Cases of grammatically licensed omission are called Constructional Null Instantiation (CNI). In the dataset examined in this study, all cases of NI are instances of CNI. There are two sources for these: genre-specific telegraphic omission of items relevant to processes being described, and grammatical omission seen with active verbs used in the passive voice, with no 'by' clause. Examples of both of these are shown below:

(8) Constructional Null Instantiation: CNI

a. genre-specific telegraphic omission

TRANSLOCATES gephyrin to submembrane microaggregates [CNI of transporting entity]

entity omitted is GRIF's target protein, *Arhgef9*

b. passive of transitive verb

CaMKII-alpha is TRANSLOCATED to the cell membranes [CNI of transporting entity]

entity omitted is not readily determined without previous knowledge of process described

Annotation Reports and Lexical Entry Reports for LUs in FN were mentioned in the previous chapter. BioFN also uses these two types of reports. Annotation Reports provide annotated examples of sentences with a target LU, ordered by valence pattern observed in the sentences. Lexical Entry Reports provide a summary of grammatical realizations of associated frame elements or their omissions, and a sorted list of valence patterns used in the examples in the Annotation Report.

In the next section, examples of BioFN Frame Reports, and a sample of BioFN annotations for a few LUs will be provided.

3.3 Case study: intracellular protein transport Frames

In chapter 1, I provided a brief description of the concept of intracellular protein transport, and a collection of texts that report on this concept. In this section I introduce two Frames that provide semantic descriptions of two separate perspectives on intracellular protein transport. I will provide a definition of the two Frames and describe their relationship to one another, and to Frames in FN. I will also present a lexicographic sample of several LUs in the Frames defined to show examples of frame semantic analysis of the texts in the collection described.

3.3.1 Protein transport Frames

The first BioFN Frame to be presented is the Protein_transport Frame. The description of this Frame, taken from its BioFN Frame Report, is as follows:

(9) Protein_transport Frame

Definition: This Frame involves the phenomenon of intracellular protein transport, the directed movement within a cell of a Transported_entity from a Transport_origin to a different location, the Transport_destination. Alternatively, Transport_locations may be mentioned with no specific indication of origin vs. destination, or the location is both origin and destination in continuous, frequent motion events. Movement of the Transported_entity follows one of a variety of transport mechanisms; these may involve crossing specific membranes, moving through pores within a subcellular component such as the nucleus, or making use of membrane-enclosed transport intermediates.

As shown below in Table 3.2, this Frame includes four core FEs. A unique color is associated with each of the FEs. BioFN provides the following definitions for these FEs:

Transported_entity		Protein or protein complex that moves from one location in a cell to another location.
Transport_origin		The location of the Transported_entity before the motion event takes place.
Transport_destination		The location of the Transported_entity after the motion event takes place.
Transport_locations		The cellular component(s) mentioned in the movement of transported entities in cases when no specific origin or destination is indicated, or the location is both origin and destination in continuous, frequent motion events.

Table 3.2: Protein_transport Core FEs

The following figure shows a BioFN annotation example using the Protein_transport Frame:

(10) inhibited TRANSLOCATION of the enzyme to the membrane

In this example, the predicator is ‘translocation’, indicated in all-capital letters with black shading. The portion of text expressing the transported entity, ‘of the enzyme’, is shaded in blue, the color chosen for this FE. Likewise, the portion of text expressing the transport destination, ‘to the membrane’, is shaded in the purple, the color designated for this FE.

The Frame Report notes that these are core FEs, and also that the location FEs form a ‘coreness set’, or ‘CoreSet’. This is a relation held across FEs in which any one of the set can satisfy a semantic valence of predicators of the Frame (Ruppenhofer et al. 2006, p.

29). In cases when two or more FEs form a CoreSet, frequently the FEs do not all occur together. The example shown above is just such a case.

No cases of non-core FEs are included in this case study, even though frequently such items might provide critical details about the event expressed by the predicator. For example, in (10), the nominal target LU is introduced as an object of the verb ‘inhibit’, together forming an expression that provides details about blocking of a transport process. While it would be useful to analyze and annotate expressions of details about the setting, technique, or consequences of protein transport events, analysis of these items is not systematically included in the data source used for this study. Analysis of non-core FEs is listed as one of the first targets for future work in the chapter 5.

Another important piece of information provided in the Frame Report for Protein_transport is that this Frame inherits from the FN Frame ‘Motion’[34]. The Frame-to-Frame relation ‘Inheritance’ specifies that facts true about the Parent Frame must correspond to an equally or more specific fact about the Child Frame (Ibid., p.105-6). The definition FN provides for the Motion Frame specifies that some entity starts out in one place and ends up in some other place, having covered some space between the two. The definition for Protein_transport is a more specific version of this, in that it specifies that the motion event takes place inside a living cell. The Protein_transport FEs ‘Transported_entity’, ‘Transport_origin’, and ‘Transport_destination’ are bound to the Motion FEs ‘Theme’, ‘Source’, and ‘Goal’, respectively.

LUs in the Protein_transport Frame include the following 32 items:

(11) Protein_transport Frame, Lexical Units

delivery.n, *efflux.n*, *endocytosis.n*, *enter.v*, *entry.n*, *exit.v*, *exocytosis.n*, *export.n*, *import.n*, *internalization.n*, *migrate.v*, *mobilization.n*, *move.v*, *movement.n*, *recruitment.n*, *recycle.v*, *recycling.n*, *redistribution.n*, *release.n*, *relocate.v*, *relocation.n*, *return.v*, *shift.n*, *shuttle.v*, *shuttling.n*, *targeting.n*, *traffic.n*, *trafficking.n*, *translocate.v*, *translocation.n*, *transport.n*, *transport.v*

The ‘.v’ or ‘.n’ included at the end of each of these items indicates its part-of-speech (POS). We see then that 22 of the items listed are nominal LUs, while the other 10 are verbal LUs.

The next BioFN Frame to be presented is the Cause_protein_transport Frame. The description of this Frame, taken from its Frame Report, is as follows:

[34] FN webpage for the Frame ‘Motion’: http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=118&frame=Motion&

(12) Cause_protein_transport Frame

Definition: A Transporting_entity, typically a protein, mediates the transport of a Transported_entity, a different molecular entity, within a cell from a Transport_origin to a Transport_destination, a different location. Alternatively, Transport_locations may be mentioned with no specific indication of origin vs. destination, or the location is both origin and destination in continuous, frequent motion events. Movement of the Transported_entity follows one of a variety of transport mechanisms; these may involve crossing specific membranes, moving through pores within a subcellular component such as the nucleus, or making use of membrane-enclosed transport intermediates.

This Frame participates in the Frame-to-Frame relation 'Causative_of' with the Protein_transport Frame. As shown in the definition, this Frame includes an agentive FE 'Transporting_entity', in addition to the four FEs that are in the Protein_transport Frame. It should be noted, causation for this activity is very complex. In many ways, the Transporting_entity *facilitates* protein transport, more than *causes* it. As suggested by Cohen (K.B. Cohen, personal communication, Aug. 5, 2009), the items denoted by entities labeled as Transporting_entity are proteins or protein complexes that either bind directly to the Transported_entity, or rather, form a vesicle which surrounds the Transported_entity, and are *causative agents* in this sense.[35] Nevertheless, the language used in these cases fits the patterns seen in other cases of causative LUs, especially the expression of an agentive entity as the subject in transitive verb phrases, uses of these transitive verbs in the passive voice, or the use of a PP with the preposition 'by' in post-nominal modifiers.

FEs common across the two Frames have the same definitions and color assignments in both Frames. The FE 'Transporting_entity' is assigned its own unique color, and provided with the definition shown in the first row of Table 3.3:

Transporting_entity		Protein that plays a critical role in mediating the transport of the transported entities.
Transported_entity		Protein or protein complex that moves from one location in a cell to another location.
Transport_origin		The location of the Transported_entity before the motion event takes place.
Transport_destination		The location of the Transported_entity after the motion event takes place.
Transport_locations		The cellular component(s) mentioned in the movement of transported entities in cases when no specific origin or destination is indicated, or the location is both origin and destination in continuous, frequent motion events.

Table 3.3: Cause_protein_transport Core FEs

As with the Protein_transport Frame, all of the FEs are core FEs, and the location FEs form a CoreSet. And again, no cases of non-core FEs for this Frame are included here.

[35] This notion of causality is similar to the concept defined in the FN Frame *Cause_motion*, a Frame which includes LUs like the verb '*throw*'.

The following figure shows an annotation example using the Cause_protein_transport Frame:

(13) Insulin receptor substrate 1 TRANSLOCATION to the nucleus by the human JC virus T-antigen

As in the previous annotation example, the lexeme of the LU is 'translocation'. The portion of text expressing the transported entity, 'Insulin receptor substrate 1', is again shaded in blue, the color chosen for this FE. The portion of text expressing the transport destination, 'to the nucleus', and the portion of text expressing the transporting entity, 'by the human JC virus T-antigen', are shaded in purple and green, respectively, the colors designated for these FEs. Again we see that expressing only one of the location FEs is licensed by the Frame, as indicated in the CoreSet status of the location FEs.

LUs in the Cause_protein_transport Frame include the following 18 items:

(14) Cause_protein_transport Frame, Lexical Units

anchor.v, *distribute.v*, *exclude.v*, *export.n*, *export.v*, *import.v*, *internalize.v*, *recruit.v*, *recycle.v*, *release.v*, *relocalize.v*, *relocate.v*, *sequester.v*, *sort.v*, *target.v*, *translocate.v*, *translocation.n*, *transport.v*

As shown in the list, the majority of the LUs for this Frame are verbal.

In 6 cases across the two Frames presented in this section, the same lexemes are used in both Frames:

(15) *export.n*, *recycle.v*, *relocate.v*, *translocate.v*, *translocation.n*, *transport.v*

Most of these cases of identical lexemes involve verbal LUs, a fact not surprising given that most of the LUs of the Cause_protein_transport are verbs. There are cases of near matches of lexemes, in which a noun vs. verb contrast can be seen across several LUs in the two Frames:

(16) Identical lexemes across two Protein_transport Frames

Protein_transport	Cause_protein_transport
import.n	import.v
release.n	release.v

A similar contrast can be seen when considering nominally vs. verbally derived lexemes in the two Frames:

(17) Derivationally related lexemes across two Protein_transport Frames

Protein_transport	Cause_protein_transport
internalization.n	internalize.v
recruitment.n	recruit.v
relocation.n	relocalize.v
targeting.n	target.v

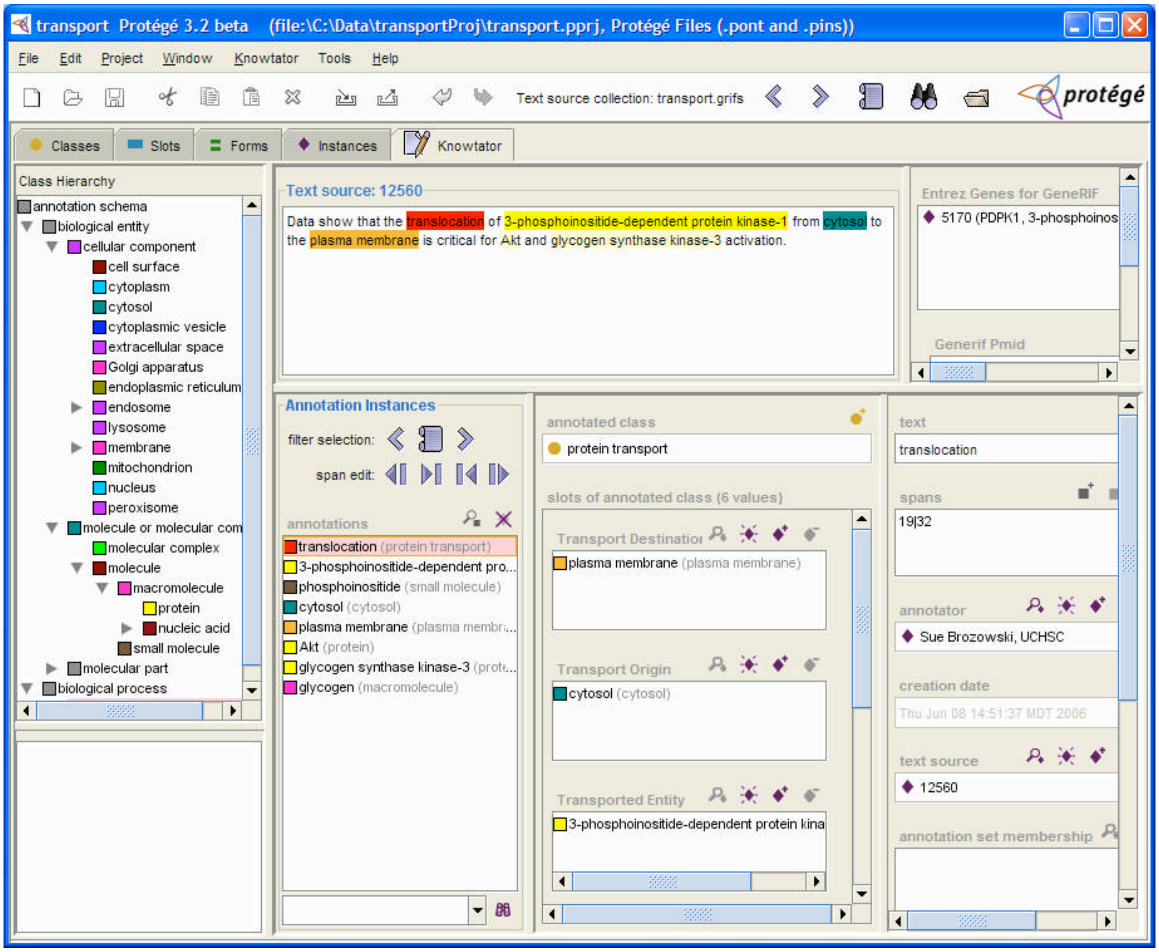
Despite these patterns, it should be noted that the number of LUs for the two Frames is not large enough to make any interesting statistical claims about choice of POS in language used for this domain. What's more, sample annotations provided later in this chapter show that assertions about protein transport events are made using both nominal and verbal LUs, from the perspective of either Frame.

In the next section, I describe protein transport GRIFs, the particular kind of molecular biology text analyzed in this study. The lexemes discussed above are used in these texts to express concepts of intracellular protein transport.

3.3.2 Transport GRIFs

As reported in chapter 1, the corpus used in this study is a collection of GRIFs gathered and analyzed by researchers in the Hunter Lab (HL). The GRIFs chosen describe and make assertions about intracellular transport of proteins. They were analyzed using Knowtator (Ogren 2006), an annotation tool integrated with the ontologically grounded knowledge base Protégé, created and maintained by members of HL. The annotations stored in the HL knowledge base (HLKB) mark key transport predicates in addition to spans of text that express core roles of transport events. These core roles are the slots of the protein transport class in the HLKB. They include entities involved in transport activities, and locations during protein transport events, usually the origin or destination. By HLKB's definition, fillers of these slots are all instances of biological entities, specifically, molecules, molecular complexes, or cellular components. The following Protégé screen shot shows the annotation of a particular GRIF in the HLKB:

(18) Screen shot of Hunter Lab's Knowledge Base


transport Protégé 3.2 beta (file:\C:\Data\transportProj\transport.pprj, Protégé Files (.pont and .pins))
File
Edit
Project
Window
Knowtator
Tools
Help
Text source collection: transport.grifs
protégé
Classes
Slots
Forms
Instances
Knowtator

Class Hierarchy

- annotation schema
- biological entity
  - cellular component
    - cell surface
    - cytoplasm
    - cytosol
    - cytoplasmic vesicle
    - extracellular space
    - Golgi apparatus
    - endoplasmic reticulum
    - endosome
    - lysosome
    - membrane
    - mitochondrion
    - nucleus
    - peroxisome
  - molecule or molecular com
    - molecular complex
    - molecule
      - macromolecule
        - protein
        - nucleic acid
      - small molecule
  - molecular part
- biological process

Text source: 12560

Data show that the translocation of 3-phosphoinositide-dependent protein kinase-1 from cytosol to the plasma membrane is critical for Akt and glycogen synthase kinase-3 activation.

Entrez Genes for GeneRIF

5170 (PDPK1, 3-phosphoinos

Generif Pmid

Annotation Instances

filter selection:

span edit:

annotations

- translocation (protein transport)
- 3-phosphoinositide-dependent pro...
- phosphoinositide (small molecule)
- cytosol (cytosol)
- plasma membrane (plasma membr...
- Akt (protein)
- glycogen synthase kinase-3 (prot...
- glycogen (macromolecule)

annotated class: protein transport

slots of annotated class (6 values)

Transport Destinatior: plasma membrane (plasma membrane)

Transport Origin: cytosol (cytosol)

Transported Entity: 3-phosphoinositide-dependent protein kina

text: translocation

spans: 19|32

annotator: Sue Brozowski, UCHSC

creation date: Thu Jun 08 14:51:37 MDT 2006

text source: 12560

annotation set membership

The following is a version of the GRIF with no annotation, and a text-only version of the annotation provided in the screenshot:

(19) Example GRIF

a. No annotation

Data show that the translocation of 3-phosphoinositide-dependent protein kinase-1 from cytosol to the plasma membrane is critical for Akt and glycogen synthase kinase-3 activation.

b. Text-only version of HLKB annotation

Data show that the translocation of [3- [phosphoinositide $_{\text{-:SmallMolecule}}$] -dependent protein kinase-1 $_{\text{TransportedEntity:Protein}}$] from [cytosol $_{\text{Origin:CellularComponent}}$] to the [plasma membrane $_{\text{Destination CellularComponent}}$] is critical for [Akt $_{\text{-:Protein}}$] and [[glycogen $_{\text{-:Macromolecule}}$] synthase kinase-3 $_{\text{-:Protein}}$] activation.

The example demonstrates that annotations in the HL GRIF collection provide four kinds of information in GRIFs:

1) a predicate that expresses an intracellular protein transport concept

And for each of the protein transport class's slot fillers that is directly realized in the GRIF:

2) colored labeling of the portion of the GRIF that is the realization of the filler
3) the name of a slot for the protein transport concept
4) the ontology class of the slot filler

In the text-only version of the screenshot, coloring is replaced with square brackets around the strings of text that express slot fillers, and in subscript notation, the name of the protein transport slot, followed by a colon, followed by the class of the filler. The example shows that other items in the GRIF that are expressions of anything of the ontology class 'biological entity' are also annotated, though since they are not expressions of protein transport slot fillers, they are not associated with any of the protein transport class's slots. These appear in the text only version as a hyphen directly before the colon. The class structure of the knowledge base and its relation to the BioFN Frames proposed here will be discussed in greater detail in the next chapter.

As illustrated in the above example, when marking items of protein transport events, only the head of the phrase of relevance is marked in the HLKB. These are the portions of a full phrase that typically appear in domain ontologies. By contrast, in annotating GRIFs, BioFN includes the full phrase of FE realizations, following the guidelines of lexicographic annotation described in the previous section. The example shown in the screenshot above looks like this in BioFN:

(20) BioFN annotation of GRIF example shown in (18)

Data show that the TRANSLOCATION of 3-phosphoinositide-dependent protein kinase-1 from cytosol to the plasma membrane is critical for Akt and glycogen synthase kinase-3 activation .

This annotation uses BioFN's definition of the Protein_transport Frame and its FEs, provided in the previous section. The most visible difference between the two annotations is the inclusion of full phrases of the FE realizations. What is missing in the BioFN annotation is specification of the domain ontology class of cellular components and named entities denoted in the text, whether part of a FE realization or not. However, the BioFN website does indicate for each GRIF which HLKB class the transport predicate has been assigned to, and what Entrez Gene IDs and PubMed document numbers are associated with the GRIF. This way it is clear for each GRIF what gene or genes are being talked about, and where more information about the genes can be found.

In the next section, I will describe how the collection of annotated GRIFs in the HLKB was modified for BioFN.

3.3.3 Transforming HLKB GRIF annotations for BioFN

Transforming annotations of the GRIFs included in this study from HLKB's version to a version for BioFN begins with locating a target predicator appearing in the GRIF that evokes a Protein_transport Frame. The source of the predicator is almost always the protein transport annotation's cover term in the HLKB. The specific BioFN Frame chosen for any particular GRIF depends on the semantics of the language of the assertion expressed in the GRIF, and in many cases, whether or not a Transporting_entity slot filler appears in the HLKB annotation for it. If such a slot filler does appear, or if it is a causal assertion expressed in the GRIF, it is annotated as an instance of the Cause_protein_transport Frame. Otherwise, it is annotated as an instance of the Protein_transport Frame.

Two common linguistic structures that suggest a causative analysis are nominal predicates in which an agentive participant is expressed with a possessive pronoun that precedes the target noun or with a prepositional phrase with 'by' following it. The following examples, in which a predicator and the proposed realization of a transporting entity FE are underlined, show this:

(21) Examples with Transporting_entity expressed

> plays a pivotal role in stimulating oxidase activity through its translocation of p47phox.
>
> results suggest a similarity in mechanism of translocation by the chaperone components HslU and ClpX

Verbal predicates are also used to express causal assertions, often in the passive voice. In these cases, though, frequently the agentive participant is not mentioned in the GRIF. As noted in section 3.2, when this sort of omission occurs, BioFN records the type of null instantiation employed and the name of the FE omitted. For passive uses of verbal predicates in which the agent is omitted, a constructional null instantiation (CNI) analysis is proposed:

(22) Transporting_entity omitted via CNI

> CaMKII-alpha is translocated to the cell membranes [CNI of Transporting_entity]

In structures with verbal predicates expressing a causative event, the verb is in the active voice and the transported entity is the direct object. The following example shows a case of this:

(23) Transporting_entity omitted via CNI; Obj expressed

translocates gephyrin to submembrane microaggregates [CNI of Transporting_entity]

Example (23) also shows a case in which the external argument of the predicate, the core FE 'Transporting_entity', is omitted. This item is actually the entity whose function is being described in the GRIF, the target gene. Cases in which there is no explicit reference to the target gene happen frequently in GRIFs. Because GRIFs are statements about particular entities, readers make the contextual inference that the unexpressed item is the entity being written about. In this example, the transporting entity was omitted using CNI.

Omission of the GRIF's target gene also happens in instances of the non-causative Protein_transport Frame, as seen in the following example:

(24) CNI omission of GRIF's target gene

activation and translocation to cell membrane dependent on protein kinase C [CNI of Transported_entity]

In example (24), the target gene of the GRIF is EG 8877, a gene whose official symbol and name are SPHK1 and sphingosine kinase 1, respectively. Neither of these two appears in the GRIF.

The process of transforming annotations from HLKB's version to a BioFN version is more straightforward when FE realizations can be based on slot fillers that are explicitly realized in the GRIF and annotated as such in the HLKB. Though I have modified the annotations so that they cover a complete constituent rather than just its head, I have otherwise tried to be faithful to the annotation decisions of the biologists hired by the Hunter Lab.[36] I have used several automated natural language processing tools to help transform the GRIF annotations, including a POS tagger for molecular biology texts created by researchers with the GENIA project[37], and a lexical parser created by the Stanford NLP Group[38]. Frequently the tagging and parsing results these tools produced required manual correction. In addition, aligning semantic analysis with syntactic parses required creating additional NLP software, along with thorough manual review and corrections. With these tasks completed, I created webpages for all of the Frame Reports and Annotation Reports, and built corresponding Lexical Entry Reports for recording and summarizing valence patterns for each LU based on frame element realizations and omissions observed in GRIFs they are used in. These Reports can be retrieved from a

[36] The members of this team frequently reviewed each other's annotation choices. I do not have the expertise required to attempt to find and correct errors in molecular biology analysis they might have missed.

[37] Webpage for the GENIA Project, with links for downloading NLP software: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi

[38] Webpage for the Stanford NLP Group, with links for downloading NLP software: http://nlp.stanford.edu/

webserver where they have been stored and organized.[39] Following the strategy used in the HLKB, for each annotated GRIF displayed, the relevant Entrez Gene ID and related PubMed document numbers are also provided.

Examples of Annotation Reports and Lexical Entry Reports for four LUs, a verbal LU and a nominal LU from each Frame, are provided in Appendices 1, 2, and 3.

3.3.4 Lexicographic sample

In this section I present a sample of BioFN annotations of GRIFs. The sample includes four different LUs, a nominal LU and a verbal LU from each Frame, and demonstrates the variety of valence patterns for LUs observed in the language of GRIFs considered in this study. The goal is to illustrate how the semantics of each Frame is expressed with nouns and with verbs, and to observe any differences in the possibilities offered by LUs from different syntactic categories. In doing so, we see how different grammatical structures realize the semantics of the Frames.

This sample covers the two basic Frames one at a time, starting with the Protein_transport Frame. Four FEs have been defined for this Frame, all of which are core. Table 3.4 provides a reminder of the names of the FEs, and the colors for text highlighting associated with them. Three of the FEs specify location within a path of transport. As pointed out in section 3.3.1, location FEs form a CoreSet and thus usually do not all co-occur in the same GRIF.

As will be seen in the summary of annotations included in this sample, BioFN differs from FN in including many examples of items that use the same valence pattern. Like FN, BioFN has as one of its goals annotating the full spectrum of valence patterns of target LUs. Multiple cases of GRIFs that use a valence pattern already recorded offer no new lexicographic information, though counts of valence patterns used could possibly reveal preferential style of expression used in writings of this domain.

Protein_transport Frame

Frame Elements	Core Type
Transport_destination	Core
Transport_locations	Core
Transport_origin	Core
Transported_entity	Core

Table 3.4: Protein_transport Core FEs

[39] As of April 2009, for users with a valid ID and password, the webserver being used can be accessed with this URL: http://dolbey.us/MediaWiki/index.php?title=Main_Page

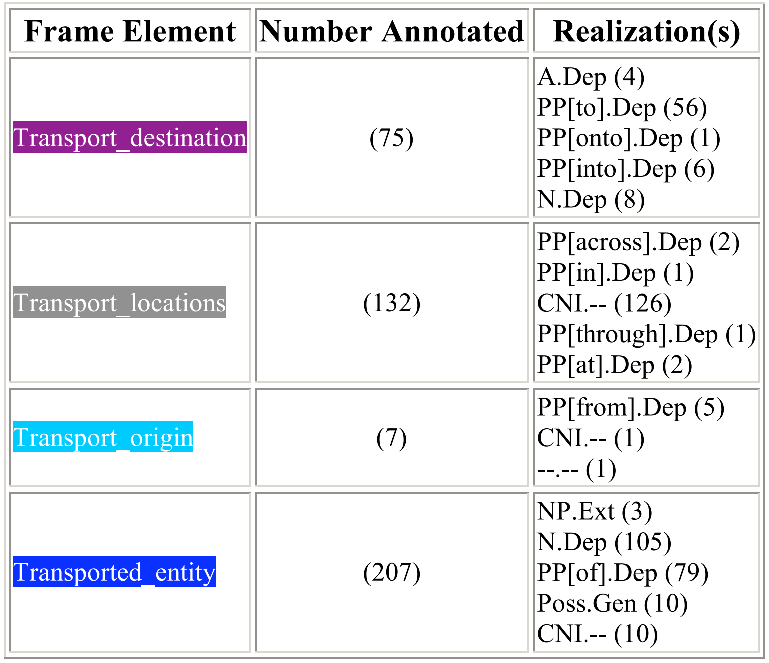
Noun target

For nominal LUs in this Frame, there are four main types of grammatical structure used in the realization of FEs:

1) PP, structure that follows the target noun
2) pre-modifier in compound with the target noun
3) possessive that precedes the target noun
4) argument of a support verb that governs the target noun; precedes target noun

In addition to these mechanisms of realization, FEs can be omitted via null instantiation.

In this section of the sample, I demonstrate examples of each of these possibilities with GRIFs that use the target LU ‘translocation.n’. There are 207 GRIFs in the collection that use this LU. Table 3.5 below, taken from its lexical entry report, shows the frequency with which FEs are realized using one of the mechanisms listed above.



Frame Element	Number Annotated	Realization(s)
Transport_destination	(75)	A.Dep (4) PP[to].Dep (56) PP[onto].Dep (1) PP[into].Dep (6) N.Dep (8)
Transport_locations	(132)	PP[across].Dep (2) PP[in].Dep (1) CNI.-- (126) PP[through].Dep (1) PP[at].Dep (2)
Transport_origin	(7)	PP[from].Dep (5) CNI.-- (1) --.-- (1)
Transported_entity	(207)	NP.Ext (3) N.Dep (105) PP[of].Dep (79) Poss.Gen (10) CNI.-- (10)

TABLE 3.5: FE realizations for annotations with *translocation.n*

In the column labeled Realization(s), GF and PT are listed with the following structure:

(25) Notation for Phrase Type and Grammatical Function

PT.GF (number of occurrences)

For cases in which the PT is PP, the particular preposition used is listed in brackets:

(26) Specifying preposition in Phrase Type notation for PPs

PP[preposition]

Null instantiation is listed with this structure:

(27) Specifying null instantiation instances

Type-of-null-instantiation.-- (number of occurrences)

These notations will also be used in the sections of this sample for the other LUs discussed.

Figure (28) shows summary counts of the four realization types listed above and of cases of null instantiation.

(28) Counts of FE realization types and null instantiation for nominal LUs

1) PP: 153
2) compound pre-modifier: 117
3) possessive: 10
4) argument of governing support verb: 3

Cases of null instantiation: 137

The sample annotations shown in this section provide examples of these strategies observed in GRIFs with this LU.

FE realizations as PPs
As seen in figure (28) above, for nominal targets, the most common grammatical structure for realizing a FE is a PP. Indeed, frequently all of the FEs realized in a GRIF with a nominal LU are done so using PPs after the target noun. The following annotations show examples of this:

(29) Data show that the TRANSLOCATION of 3-phosphoinositide-dependent protein kinase-1 from cytosol to the plasma membrane is critical for Akt and glycogen synthase kinase-3 activation .

Data show that the TRANSLOCATION [$_1$ of 3-phosphoinositide-dependent protein kinase-1] [$_2$ from cytosol] [$_3$ to the plasma membrane] is critical for Akt and glycogen synthase kinase-3 activation .

1		2	3
Transported_entity	Target LU	Transport_origin	Transport_destination
PP[of].Dep	translocation.n	PP[from].Dep	PP[to].Dep

Target gene: EG 5170, Hs: PDPK1; 3-phosphoinositide dependent protein kinase-1[40]

(30) Neuregulin stimulation causes TRANSLOCATION of ErbB4 into lipid rafts and these are necessary for signaling by ErbB4

Neuregulin stimulation causes TRANSLOCATION [$_1$ of ErbB4] [$_2$ into lipid rafts] and these are necessary for signaling by ErbB4[41]

	1	2
Target LU	Transported_entity	Transport_destination
translocation.n	PP[of].Dep	PP[into].Dep

Target gene: EG 59323, Rn: Erbb4; v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)

The order of FE realizations that are PPs varies. In the previous example, the realization of the transported entity FE comes directly after the verb, whereas in the following example, it is the realization of a location FE that comes in this position. Generally, if more than one location FE is realized with PPs in a GRIF they will be directly adjacent to one another.

(31) increased TRANSLOCATION from the cytoplasm to the membrane of protein kinase theta , a T cell signaling molecule that colocalizes with the TCR within the supramolecular activation cluster

increased TRANSLOCATION [$_1$ from the cytoplasm] [$_2$ to the membrane] [$_3$ of protein kinase theta , a T cell signaling molecule that colocalizes with the TCR within the supramolecular activation cluster]

	1	2	3
Target LU	Transport_origin	Transport_destination	Transported_entity
translocation.n	PP[from].Dep	PP[to].Dep	PP[of].Dep

Target gene: EG 5588, Hs: PRKCQ; protein kinase C, theta

In one of the GRIFs with this LU, a constituent that directly realizes one FE also allows inferences about a second FE, a frame semantic possibility FN calls *FE conflation*

[40] When showing GRIF examples, the NCBI identification of the specific gene associated with the GRIF will be displayed using the notation format shown in example (1). In this notation, first the Entrez Gene (EG) number is given, then the associated official gene symbol, followed by a semicolon and then the official name or a recorded alias: *EG number: symbol; name/alias*. The name or alias sometimes includes commas.

[41] The lexeme 'cause' used in this GRIF desctibes a relationship between the process of protein stimlation and the transport of a different protein. The protein involved in the stimulation process (*Neuregulin*) does not bind with the protein that is transported (*ErbB4*), and thus is not tagged as Transporting_entity. This sort of description shows up in seveal other GRIFs in the collection analyzed in this study.

(Ruppenhofer et al. 2006). The following example shows the annotation of this case. Here the transported FE ‘of intracellular GLUT1 transporters’ includes material that allows inferences about the Transport_origin FE, namely that the origin of transport is somewhere within the cell. In its annotation, the inferred FE is only assigned a FE label, not a GF or a PT.

(32) IL-3 caused TRANSLOCATION of intracellular GLUT1 transporters to the cell surface

IL-3 caused TRANSLOCATION [$_1$ of [$_2$ intracellular] GLUT1 transporters] [$_3$ to the cell surface]

	1	2	3
Target LU	Transported_entity	Transport_origin	Transport_destination
translocation.n	PP[of].Dep	--.--	PP[to].Dep

Target gene: EG 20525, Mm: Slc2a1; solute carrier family 2 (facilitated glucose transporter), member 1

Target gene also known as ‘Glut1’

As reported in the FE definitions for Transport Frames in section 3.3.1, the Transport_locations FE refers to cellular component(s) mentioned in the movement of transported entities in cases when no specific origin or destination is indicated, or the location is both origin and destination in continuous, frequent motion events. The following example shows a GRIF in which the PP ‘at the PM’ denotes exactly this sort of location specification.[42]

(33) ERalpha requires serine 522 for efficient TRANSLOCATION and function of ERalpha at the PM

ERalpha requires serine 522 for efficient TRANSLOCATION and function [$_1$ of ERalpha] [$_2$ at the PM]

	1	2
Target LU	Transported_entity	Transport_locations
translocation.n	PP[of].Dep	PP[at].Dep

Target gene: EG 13982, Mm: Esr1; estrogen receptor 1 (alpha)

Target gene also known as ‘ERalpha’

FE realizations as compound pre-modifier

FEs are frequently realized as a pre-modifier of the target nominal LU in a noun compound. This is the second most frequently used grammatical structure seen in GRIFs with ‘translocation.n’. The following annotation shows an example in which the transported entity FE is realized as a bare noun in a compound pre-modifier.

[42] PM is a commonly used abbreviation for the cellular component ‘plasma membrane’.

(34) IRS-3 expression blocked glucose/IGF-1 induced IRS-2 TRANSLOCATION from the cytosol to the plasma membrane , dampening IRS-2/IGF-1R interaction and subsequent activation of the PI3K/PKB/GSK3 signaling pathway

IRS-3 expression blocked [$_1$ glucose/IGF-1 induced IRS-2] TRANSLOCATION [$_2$ from the cytosol] [$_3$ to the plasma membrane] , dampening IRS-2/IGF-1R interaction and subsequent activation of the PI3K/PKB/GSK3 signaling pathway

1		2	3
Transported_entity	Target LU	Transport_origin	Transport_destination
N.Dep	translocation.n	PP[from].Dep	PP[to].Dep

Target gene: EG 84021, Rn: Irs3; insulin receptor substrate 3

Compound structures like the one seen above are given a detailed analysis in Rosario and Hearst (2004, 2005), and Nakov and Hearst (2006). In these works, the authors present methods for characterizing key semantic relations between nouns in noun-noun compounds, information which can then be use for NLP tasks like information extraction.

Frequently in GRIFs in this collection the nominal LU is part of a coordination structure, and in these cases, FE realizations are not necessarily adjacent to the target noun. This is seen in the following example.

(35) HNE causes protein kinase C (PKC) activation and TRANSLOCATION from cytosol to plasma membrane , required for HNE-induced ROS generation and other responses

HNE causes [$_1$ protein kinase C (PKC)] activation and TRANSLOCATION [$_2$ from cytosol] [$_3$ to plasma membrane] , required for HNE-induced ROS generation and other responses

1		2	3
Transported_entity	Target LU	Transport_origin	Transport_destination
N.Dep	translocation.n	PP[from].Dep	PP[to].Dep

Target gene: EG 1991, Hs: ELANE; elastase, neutrophil expressed

Target gene also known as 'HNE'

Location FEs are also sometimes realized as a bare noun pre-modifier of the target noun, as seen with the transport destination FE in the following example.

(36) two fatty acids inhibited the phorbol 12-myristate 13-acetate (PMA)-induced plasma membrane TRANSLOCATION of protein kinase C (PKC)-alpha and -epsilon

two fatty acids inhibited the phorbol 12-myristate 13-acetate (PMA)-induced [$_1$ plasma membrane] TRANSLOCATION [$_2$ of protein kinase C (PKC)-alpha and –epsilon]

1		2
Transport_destination	Target LU	Transported_entity
N.Dep	translocation.n	PP[of].Dep

Target gene 1: EG 5578, Hs: PRKCA; protein kinase C, alpha

Target gene1 also known as 'PKC-alpha'

Target gene 2: EG 5581, Hs: PRKCE; protein kinase C, epsilon

Target gene 2 also known as 'nPKC-epsilon'

Locations in GRIFs can also be expressed as adjectival pre-modifiers, as seen in the following examples.

(37) TNF binding induces release of AIP1 (DAB2IP) from TNFR1 , resulting in cytoplasmic TRANSLOCATION and concomitant formation of an intracellular signaling complex comprised of TRADD , RIP1 , TRAF2 , and AIPl .

TNF binding induces release of AIP1 (DAB2IP) from TNFR1 , resulting in [$_1$ cytoplasmic] TRANSLOCATION and concomitant formation [$_2$ of an intracellular signaling complex comprised of TRADD , RIP1 , TRAF2 , and AIPl] .

1		2
Transport_destination	Target LU	Transported_entity
A.Dep	translocation.n	PP[of].Dep

Target gene 1: EG 280943, Bt: official symbol not given; full name not given
Target gene 1 also known as 'TNF'
Target gene 2: EG 282527, Bt: official symbol not given; full name not given
Target gene 2 also known as 'TNFRSF1A'
Target gene 3: EG 504707, Bt: official symbol not given; full name not given
Target gene 3 also known as 'TRADD'

(38) recoverin *undergoes* light-dependent intracellular TRANSLOCATION in mouse rod photoreceptors

[$_1$ recoverin] *undergoes* light-dependent [$_2$ intracellular] TRANSLOCATION in mouse rod photoreceptors

1		2	
Transported_entity	Support-Verb	Transport_destination	Target LU
NP.Ext	undergo	A.Dep	translocation.n

Target gene: EG 19674, Mm: Rcvrn; recoverin

External realization of a FE with the support verb 'undergo' will be discussed later in this section.

In some GRIFs, two separate FEs are both realized as adjacent pre-nominal modifiers. The following example shows a case of this. Here the first FE expressed is the transported entity, followed by expression of the transport destination FE.

(39) The Y221 site in transfected rat CrkII regulates Rac membrane TRANSLOCATION upon cell adhesion , which is necessary for activation of downstream Rac signaling pathways .

The Y221 site in transfected rat CrkII regulates [$_1$ Rac] [$_2$ membrane] TRANSLOCATION upon cell adhesion , which is necessary for activation of downstream Rac signaling pathways .

1	2	
Transported_entity	Transport_destination	Target LU
N.Dep	A.Dep	translocation.n

Target gene: EG 54245, Rn: Crk; v-crk sarcoma virus CT10 oncogene homolog (avian)

Analysis of the structure seen in (39), in which two different FEs are realized as separate pre-nominal modifiers, could be a useful addition for tools used for automatic semantic role labeling. This idea will be discussed in the section on future work in Chapter 5.

FE realizations as possessive phrase

In some GRIFs the transported entity FE is realized as a possessive pronoun that governs the target noun. An example of this is shown below.

(40) SCD1 deficiency specifically increases CTP:choline cytidylyltransferase activity by promoting its TRANSLOCATION into membrane and enhances phosphatidylcholine biosynthesis in liver

SCD1 deficiency specifically increases CTP:choline cytidylyltransferase activity by promoting [$_1$ its] TRANSLOCATION [$_2$ into membrane] and enhances phosphatidylcholine biosynthesis in liver

1		2
Transported_entity	Target LU	Transport_destination
Poss.Gen	translocation.n	PP[into].Dep

Target gene: EG 20249, Mm: Scd1; stearoyl-Coenzyme A desaturase 1
Target gene: EG 13026, Mm: Pcyt1a; phosphate cytidylyltransferase 1, choline, alpha isoform

The possessive pronoun expressing the transported entity can also govern other activity nouns in addition to the target noun 'translocation'. In cases like this, the possessive pronoun might not be adjacent to the target noun, as seen in the following example.

(41) HSP25 binds to protein kinase C delta to inhibit its kinase activity and TRANSLOCATION to the membrane , which results in reduced cell death .

HSP25 binds to protein kinase C delta to inhibit [$_1$ its] kinase activity and TRANSLOCATION [$_2$ to the membrane] , which results in reduced cell death .

1		2
Transported_entity	Target LU	Transport_destination
Poss.Gen	Translocation.n	PP[to].Dep

Target gene: EG 15507, Mm: Hspb1; heat shock protein 1
Target gene also known as 'Hsp25'

FE realizations as arguments of governing support verbs

A common grammatical structure for nominal LUs is the use of a support verb that governs the target noun. FEs realized as arguments of a support verb are given the GF Ext, the abbreviation for external argument. Often the support verb evokes a Frame separate from the Protein_transport Frame. Though the support verb does not introduce protein transport semantics of its own to the expression, it does provide additional grammatical argument slots which can be used for the realization of FEs of the nominal Protein_transport LU. Content of the FE realization in these cases is shared between the Frame of the support verb and the Protein_transport Frame. In addressing this sort of argument sharing between semantically unrelated Frames, FN has proposed introducing a category of special governors for nominal LUs called Concomitant (Ruppenhofer 2006).

This category is intended to cover cases in which the support verb evokes a Frame that is related via a background scenario.[43] The following example shows a case of this. In this example, ‘require’, a verb that evokes the Have_as_requirement Frame, is the support verb for ‘translocation’. ‘GLUT4’, the subject of the verb ‘require’, expresses the Required_entity FE of Have_as_requirement as well as the Transported_entity FE of Protein_transport. In all of the cases of nominal LUs observed in this study, FEs that are realized as external arguments of a support verb are FEs of entities, most often the transported entity. In the collection considered here, Transport locations are never realized this way.

(42) These results suggest that GLUT4 *requires* TRANSLOCATION to the plasma membrane , as well as activation at the plasma membrane , to initiate glucose uptake , and both of these steps normally require PI 3-kinase activation .

These results suggest that [$_1$ GLUT4] *requires* TRANSLOCATION [$_2$ to the plasma membrane] , as well as activation at the plasma membrane , to initiate glucose uptake , and both of these steps normally require PI 3-kinase activation .

1			2
Transported_entity	Support-Verb	Target LU	Transport_destination
NP.Ext	require	translocation.n	PP[to].Dep

Target gene: EG 20528, Mm: Slc2a4; solute carrier family 2 (facilitated glucose transporter), member 4

Target gene also known as ‘Glut4’

Null instantiation of FEs

As noted at the beginning of this chapter, in the collection of GRIFs considered in this study, the transported entity is sometimes omitted via constructional null instantiation (CNI), a kind of omission characteristic of the telegraphic style of this genre. This is marked in annotations by including at the end of the text for the GRIF the tag ‘CNI’ in the color assigned to the FE that is omitted. The following annotation shows an example of this.

(43) activation and TRANSLOCATION to cell membrane dependent on protein kinase C CNI

activation and TRANSLOCATION [$_1$ to cell membrane] dependent on protein kinase C

		1
Transported_entity	Target LU	Transport_destination
CNI	translocation.n	PP[to].Dep

Target gene: EG 8877, Hs: SPHK1; sphingosine kinase 1

As noted in section 3.2, occasionally the transported entity does appear in the GRIF out of the grammatically local context of the target LU, and has been marked as the transported entity in HLKB. In these cases, the annotations in BioFN specify the relevant text by formatting it in italics font in the color assigned to the FE. The following

[43] FrameNet refers to this sort of support verb as a Controller (Ruppenhofer et al. 2006). Controllers are distinct from support verbs that offer no additional semantics of their own, but rather only project arguments for the governed noun, e.g. ‘make’ in “make a statement”. In this work I use the term **support verb** for both types since I do not need to distinguish the two kinds.

example demonstrates an example of this annotation style. ‘Smo’ has been indicated in HLKB to be the transported entity, and has been formatted here in italics and in the color assigned to this FE. As indicated in (44), the GRIF displayed is associated with the Entrez Gene entity ‘Smo’. However, this assertion is not guaranteed by the grammar of the text of the GRIF, and is in fact defeasible. For this reason, it is annotated as CNI.

(44) Hh-dependent TRANSLOCATION to cilia is essential for *Smo* activity , suggesting that *Smo* acts at the primary cilium CNI

Hh-dependent TRANSLOCATION [$_1$ to cilia] is essential for *Smo* activity , suggesting that Smo acts at the primary cilium

		1
Transported_entity	Target LU	Transport_destination
CNI	translocation.n	PP[to].Dep

Target gene: EG 319757, Mm: Smo; smoothened homolog (Drosophila)

In some GRIFs, there is no mention of any particular location of transport. For cases like these, BioFN associates CNI with the Transport_locations FE, since the definition for this FE already leaves the notion of direction unspecified. The following annotation shows such an example.

(45) Mutations in the tyrosine kinase domain of the IGF-IR abrogate TRANSLOCATION of the IRS-2 and IRS-2 proteins . CNI

Mutations in the tyrosine kinase domain of the IGF-IR abrogate TRANSLOCATION [$_1$ of the IRS-2 and IRS-2 proteins] .

1		
Transported_entity	Target LU	Transport_locations
PP[of].Dep	translocation.n	CNI

Target gene 1: EG 16367, Mm: Irs1; insulin receptor substrate 1
Target gene 2: EG 16001, Mm: Igf1r; insulin-like growth factor I receptor
Target gene 3: EG 384783, Mm: Irs2; insulin receptor substrate 2

As with cases of transported entities that are mentioned out of the grammatically local context of the target LU, if a relevant location is mentioned in the GRIF, and has been specified as such in HLKB, the location is formatted in italics and is placed in the color assigned to the appropriate FE. This is shown in the following example.

(46) TRANSLOCATION of Smac along with cytochrome c and other mitochondrial pro-apoptotic proteins represent important regulatory checkpoints for *mitochondria*-mediated apoptosis CNI

TRANSLOCATION [$_1$ of Smac along with cytochrome c and other mitochondrial pro-apoptotic proteins] represent important regulatory checkpoints for mitochondria-mediated apoptosis

	1	
Target LU	Transported_entity	Transport_locations
translocation.n	PP[of].Dep	CNI

Target gene: EG 56616, Hs: DIABLO; diablo homolog (Drosophila)

Target gene also known as 'SMAC'

There are GRIFs in the collection in which none of the FEs has an explicit realization. For these cases, an analysis of CNI will be specified for Transported_entity and for Transport_locations, as shown in the following example.

(47) *PKC-alpha* shows variable patterns of TRANSLOCATION in response to different stimulatory agents . CNI CNI

PKC-alpha shows variable patterns of TRANSLOCATION in response to different stimulatory agents .

Target LU	Transported_entity	Transport_locations
translocation.n	CNI	CNI

Target gene: EG 24680, Rn: Prkca; protein kinase C, alpha

Verb target

For verbal LUs, there are three main types of grammatical structure used in the realization of FEs:

1) PP, structure that follows the target verb
2) External arg., subject of target verb or verb that governs target verb
3) External arg., antecedent of relative clause in which target verb is embedded

In addition to these mechanisms of realization, as with nominal LUs, FEs can be omitted via null instantiation.

In this section of the sample, I show examples of each of these possibilities with GRIFs that use the verbal target LU 'translocate.v' in the Protein_transport Frame. There are 10 GRIFs in the collection that use this LU. The table below, taken from its lexical entry report, shows the frequency with which FEs of the Protein_transport Frame are realized using one of the mechanisms listed above.

Frame Element	Number Annotated	Realization(s)
Transport_destination	(8)	PP[to].Dep (8)
Transport_locations	(2)	PP[across].Dep (1) CNI.-- (1)
Transport_origin	(3)	PP[from].Dep (2) CNI.-- (1)
Transported_entity	(10)	NP.Ext (9) CNI.-- (1)

TABLE 3.6: FE realizations for annotations with *translocate.v*

Location FE realizations as PPs, Transported entity FE as External NP
In the case of verbal LUs, the only FEs realized as PPs are location FEs. Transported entity items are never expressed with PPs, but are expressed instead as external NPs, either the subject of the target verb or the subject of a verb that governs the target verb. The following annotations show examples of this.

(48) CERK TRANSLOCATES during activation from the cytosol to a lipid raft fraction

[$_1$ CERK] TRANSLOCATES during activation [$_2$ from the cytosol] [$_3$ to a lipid raft fraction]

1		2	3
Transported_entity	Target LU	Transport_origin	Transport_destination
NP.Ext	translocate.v	PP[from].Dep	PP[to].Dep

Target gene: EG 64781, Hs: CERK; ceramide kinase

(49) Nm23-H2 had *a cytoplasmic and nuclear localization* but was *induced* to TRANSLOCATE to the plasma membrane upon stimulation of thromboxane A2 receptor beta to show extensive co-localization with the receptor . CNI

Nm23-H2 had *a cytoplasmic and nuclear localization* but was *induced* to TRANSLOCATE [$_1$ to the plasma membrane] upon stimulation of thromboxane A2 receptor beta to show extensive co-localization with the receptor .

1			2
Transported_entity	Transport_origin	Target LU	Transport_destination
NP.Ext	CNI	translocate.v	PP[to].Dep

Target gene 1: EG 6915, Hs: TBXA2R; thromboxane A2 receptor
Target gene 2: EG 4831, Hs: NME2; non-metastatic cells 2, protein (NM23B) expressed in
Target gene 2 also known as 'NM23-H2'

(50) hic-5 is a mediator of tensional force , TRANSLOCATING directly from focal adhesions to actin stress fibers upon mechanical stress and regulating the contractile capability of cells in the stress fibers

[$_1$ hic-5] is a mediator of tensional force , TRANSLOCATING directly [$_2$ from focal adhesions] [3 to actin stress fibers] upon mechanical stress and regulating the contractile capability of cells in the stress fibers

1		2	3
Transported_entity	Target LU	Transport_origin	Transport_destination
NP.Ext	translocate.v	PP[from].Dep	PP[to].Dep

Target gene: EG 21804, Mm: Tgfb1i1; transforming growth factor beta 1 induced transcript 1
Target gene also known as 'hic-5'

In several GRIFs in this collection, the transported entity FE is expressed as an NP that is modified by a relative clause in which the target verb is embedded. BioFN follows FN's lead and annotates as transported entity both the NP that denotes the entity and the relative pronoun that introduces the clause that contains the target verb. An example of this is shown in the following annotation.

(51) proposal that STIM1 functions as the missing link between Ca2+ store depletion and store-operated calcium influx , serving as a Ca2+ sensor that TRANSLOCATES upon store depletion to the plasma membrane to activate CRAC channels

proposal that STIM1 functions as the missing link between Ca2+ store depletion and store-operated calcium influx , serving as [$_1$ a Ca2+ sensor] that TRANSLOCATES upon store depletion [$_2$ to the plasma membrane] to activate CRAC channels

1		2
Transported_entity	Target LU	Transport_destination
NP.Ext	translocate.v	PP[to].Dep

Target gene: EG 117086 (replaced with 361618), Rn: Stim1; stromal interaction molecule 1
Target gene: EG 32556, Dm: Stim; Stromal interaction molecule

Null instantiation of FEs

Cases of null instantiation also happen with verbal LUs of the Protein_transport Frame in this collection. Both location FEs and transported entity FEs are occasionally omitted via CNI. Annotation for these cases is the same as it is for nominal LUs, as seen in the following example. The PP after the target verb here, "to the vicinity of the immune synapse after T cell receptor stimulation", was not annotated as a filler for the transport destination slot in HLKB because the phrase does not denote a well-defined cellular component. I have opted not to alter HL's annotation for this, but instead analyze the transport locations as a case of CNI.

(52) By rapidly TRANSLOCATING to the vicinity of the immune synapse after T cell receptor stimulation , *Pyk2* plays an essential role in T cell activation and polarized secretion of cytokines . CNI CNI

By rapidly TRANSLOCATING to the vicinity of the immune synapse after T cell receptor stimulation , Pyk2 plays an essential role in T cell activation and polarized secretion of cytokines .

Target LU	Transport_destination	Transported_entity
translocate.v	CNI	CNI

Target gene: EG 2185, Hs:PTK2B; PTK2B protein tyrosine kinase 2 beta

Target gene also known as 'PYK2'

Cause_protein_transport Frame

This portion of the sample covers the Cause_protein_transport Frame, which has a causative relation with the previous Frame covered in the sample. Five FEs have been defined for this Frame, all of which are core. The four FEs of the Protein_transport Frame (three location FEs and a FE for the transported entity) are also included in the definition of Cause_protein_transport, and are associated with the same color assignments. As with the Protein_transport Frame, the three location FEs of this Frame form a CoreSet. In addition to these FEs, there is a FE in this Frame for Transporting_entity, a protein that either binds directly to the Transported_entity, or rather, forms a vesicle which surrounds the Transported_entity, and thus plays a critical role in facilitating the transport event.

The following table provides a reminder of the names of the FEs, and the colors for text highlighting associated with them.

Frame Elements	Core Type
Transport_destination	Core
Transport_locations	Core
Transport_origin	Core
Transported_entity	Core
Transporting_entity	Core

Table 3.7: Cause_protein_transport Core FEs

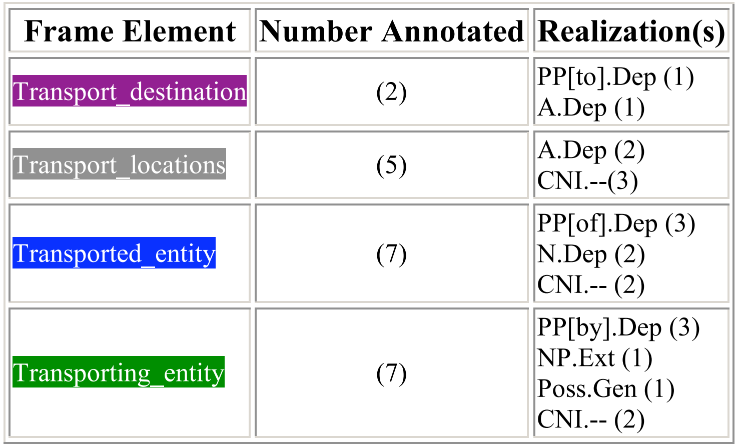
Noun target

The same four types of grammatical structure that can be used for realizing FEs with nominal LUs in the Protein_transport Frame are also available in the Cause_protein_transport Frame. They are listed again below for convenience. Having an additional FE that is sometimes explicitly realized does not affect the grammatical structure observed in the GRIFs in the collection. In each GRIF with a nominal LU of this Frame, between one and three FEs are expressed, the same as is the case for nominal LUs of the Protein_transport Frame.

1) PP, structure that follows the target noun
2) pre-modifier in compound with the target noun
3) possessive that precedes the target noun
4) argument of a support verb that governs the target noun; precedes target noun

Again, in addition to these mechanisms of realization, FEs can be omitted via null instantiation.

In this section of the sample, I demonstrate examples of each of these possibilities with GRIFs that use the Cause_protein_transport target LU ‘translocation.n’, a LU with the same lexeme as the nominal LU from the Protein_transport Frame discussed earlier in the sample. There are only 7 GRIFs in the collection that use this LU, far fewer than with Protein_transport.translocation.n.[44] The table below, taken from its lexical entry report, shows the frequency with which FEs are realized using one of these mechanisms listed above.



Frame Element	Number Annotated	Realization(s)
Transport_destination	(2)	PP[to].Dep (1) A.Dep (1)
Transport_locations	(5)	A.Dep (2) CNI.--(3)
Transported_entity	(7)	PP[of].Dep (3) N.Dep (2) CNI.-- (2)
Transporting_entity	(7)	PP[by].Dep (3) NP.Ext (1) Poss.Gen (1) CNI.-- (2)

TABLE 3.8: FE realizations for annotations with *Cause_protein_transport.translocation.n*

Figure (53) shows summary counts of the four realization types listed above and of cases of null instantiation.

[44] In order to avoid ambiguity and confusion in cases in which the same lexeme is used for LUs of different Frames, this fuller notation is sometimes used to identify the LU: Frame.lexeme.POS

(53) Counts of FE realization types and null instantiation for nominal LUs

1) PP: 7
2) compound pre-modifier: 5
3) possessive: 1
4) argument of governing support predicate: 1

Cases of null instantiation: 7

These types of realization are the same as those seen with the nominal LU of the Protein_transport Frame. And again, the most frequent types of FE realization are PPs and compound pre-modifier of the target nominal. The sample annotations shown in this section provide examples of these strategies observed in GRIFs with this LU. The focus will be on realizations of the agentive FE of this Frame, Transporting_entity. This FE is realized in each of the four structures listed above except for the second one, compound pre-modifier. Transporting_entity is not realized as a compound pre-modifier with a target nominal LU in any of the GRIFs considered in this study.

Transporting_entity FE realization as PP

In the following examples, the transporting entity FE is realized with a PP in which the preposition is ‘by’. In the first example, the transported entity is realized as a PP, and transport location is omitted via CNI. In the second example, the transported entity is realized as a pre-modifier of the target noun and the destination is realized as a PP.

(54) role of PRC1 in midzone formation , indicate that cell cycle-dependent TRANSLOCATION of PRC1 by Kif4 is essential for midzone formation and cytokinesis . CNI

role of PRC1 in midzone formation , indicate that cell cycle-dependent TRANSLOCATION [$_1$ of PRC1] [$_2$ by Kif4] is essential for midzone formation and cytokinesis .

	1	2	
Target LU	Transported_entity	Transporting_entity	Transport_locations
translocation.n	PP[of].Dep	PP[by].Dep	CNI

Target gene 1: EG 24137, Hs: KIF4A; kinesin family member 4A

Target gene 1 also known as ‘KIF4’

Target gene 2: EG 9055, Hs: PRC1; protein regulator of cytokinesis 1

(55) Insulin receptor substrate 1 TRANSLOCATION to the nucleus by the human JC virus T-antigen

[$_1$ Insulin receptor substrate 1] TRANSLOCATION [$_2$ to the nucleus] [$_3$ by the human JC virus T-antigen]

1		2	3
Transported_entity	Target LU	Transport_destination	Transporting_entity
N.Dep	translocation.n	PP[to].Dep	PP[by].Dep

Target gene 1: EG 16367, Mm: Irs1; insulin receptor substrate 1

Target gene 2: EG 3667, Hs: IRS1; insulin receptor substrate 1

Realization of the Transporting_entity as a PP with ‘by’ in these cases provides evidence that the GRIFs are instances of the Cause_protein_transport Frame.

Transporting_entity FE realization as possessive noun or pronoun

Transporting_entity can also be realized as a possessive noun or pronoun that governs the target noun, as seen in the following example. The example also shows another instance in which the target noun is part of a coordination phrase, and follows a noun expressing an additional activity the transporting entity plays a role in.

(56) Protein kinase C delta plays a pivotal role in stimulating monocyte NADPH oxidase activity through its regulation of phosphorylation and TRANSLOCATION of p47phox . CNI

Protein kinase C delta plays a pivotal role in stimulating monocyte NADPH oxidase activity through [$_1$ its] regulation of phosphorylation and TRANSLOCATION [$_2$ of p47phox].

1		2	
Transporting_entity	Target LU	Transported_entity	Transport_locations
Poss.Gen	translocation.n	PP[of].Dep	CNI

Target gene: EG 5580, Hs: PRKCD; protein kinase C, delta

In a phrase like ‘its translocation’, the constituent ‘its’ is ambiguous in that it could be an expression of one of two different FEs in the Cause_protein_transport Frame, either the Transporting_entity or the Transported_entity, a within-Frame ambiguity. Another interpretation is that it could be an expression of the Transported_entity FE of the Protein_transport Frame, a cross-Frame ambiguity. The presence of the post-nominal PP expressing the transported entity ‘of p47 phox’ eliminates the potential ambiguity.

Transporting_entity FE realization as argument of governing support predicate

In addition to the case of a support verb governing a nominal LU, there can be other types of support predicates that govern a target noun. In fact, support expressions can include several layers of grammatical structures in which interpretation of non-local noun phrases is guaranteed by the grammar of the governing structures. The following GRIF shows an example of this sort of embedded non-local realization of the Transporting_entity FE. In this case, the pronoun ‘it’ is the realization of this FE. The target noun is governed by the support noun ‘carrier’, which in turn, is governed by the support verb ‘act’. The Transporting_entity FE for ‘translocation’ is realized as the subject of ‘act’. Though this is a more complex embedding of grammatical structures and dependency relations than seen in previous instances of syntactic governing in this chapter, the overall structure nevertheless guarantees the interpretation of the non-local NP. The other relevant FEs, Transport_locations and Transported_entity’, are realized as local constituents of the target noun, either a compound pre-modifier or a post-nominal PP.

(57) novel aspect of RXRalpha function : it *acts* as a *carrier* for nucleocytoplasmic TRANSLOCATION of orphan receptors

novel aspect of RXRalpha function : [$_1$ it] *acts* as a *carrier* for [$_2$ nucleocytoplasmic] TRANSLOCATION [$_3$ of orphan receptors]

1		2		3
Transporting_entity	Support Preds	Transport_locations	Target LU	Transported_entity
NP.Ext	act.v; carrier.n	N.Dep	translocation.n	PP[of].Dep

Target gene: EG 6256, Hs: RXRA; retinoid X receptor, alpha

Null instantiation of Transporting_entity FE

In GRIFs with LUs of the Cause_protein_transport Frame, the Transporting_entity FE is frequently omitted via CNI. Annotation for these omissions is done the same here as it is with FE omissions in the Protein_transport Frame: the tag ‘CNI’ in the color assigned to the omitted FE is included at the end of the text for the GRIF. The following annotation shows an example of this.

(58) polyamine depletion increased expression of the NPM gene and enhances NPM nuclear TRANSLOCATION and increased NPM interacts with and stabilizes p53 , leading to inhibition of IEC-6 cell proliferation CNI

polyamine depletion increased expression of the NPM gene and enhances [$_1$ NPM] [$_2$ nuclear] TRANSLOCATION and increased NPM interacts with and stabilizes p53 , leading to inhibition of IEC-6 cell proliferation

1	2		
Transported_entity	Transport_destination	Target LU	Transporting_entity
N.Dep	A.Dep	translocation.n	CNI

Target gene: EG 4869, Hs: NPM1; nucleophosmin (nucleolar phosphoprotein B23, numatrin)

Target gene also known as ‘NPM’

As seen previously in this sample, there are two sorts of omission that occur in this collection. In some cases, the GRIF itself might have no mention at all of a transporting entity, even in portions of the GRIF not local to the target LU. The previous example shows a case of this. In other cases, the transporting entity is mentioned in the GRIF though not local to the target. When the transporting entity in these cases is marked in the HLKB, the relevant portion of text is italicized and placed in the color appropriate for the FE. The following example shows this form of annotation. Here there is CNI of both the transported entity and the transporting entity, with mention of only the latter.

(59) Full binding of androgen to the polyglutamine-expanded N-terminal domain of *the mutant AR* leads to structural alteration with nuclear TRANSLOCATION that eventually results in the onset of spinal and bulbar muscular atrophy . CNI CNI

Full binding of androgen to the polyglutamine-expanded N-terminal domain of the mutant AR leads to structural alteration with [$_1$ nuclear] TRANSLOCATION that eventually results in the onset of spinal and bulbar muscular atrophy .

	1		
Transporting_entity	Transport_destination	Target LU	Transported_entity
CNI	A.Dep	translocation.n	CNI

Target gene: EG 367, Hs: AR; androgen receptor

Verb target

For verbal LUs, there are three main types of grammatical structure used in the realization of FEs:

1) PP, structure that follows the target verb
2) Obj, direct object of target verb
3) External arg., subject of target verb or verb that governs target verb
4) External arg., antecedent of relative clause in which target verb is embedded

For verbal LUs in the Cause_protein_transport Frame, the Transporting_entity is most often omitted via null instantiation, usually CNI in passive verb structures, but occasionally telegraphic CNI associated with this genre. In this section of the sample, I will show examples of each type.

GRIFs with the verbal LU Cause_protein_transport.translocate.v show many cases of CNI. Two of these are shown below. In both cases, the target verb is in the passive voice, and the Transported_entity FE is expressed as its subject, an external NP. Location FEs are realized as post-verbal PPs.

(60) In response to an increase of cellular cholesterol , fatty acid transporter is TRANSLOCATED from cytosol to membranes of type II pneumocytes . CNI

In response to an increase of cellular cholesterol , [$_1$ fatty acid transporter] is TRANSLOCATED [$_2$ from cytosol] [$_3$ to membranes of type II pneumocytes] .

1		2	3	
Transported_entity	Target LU	Transport_origin	Transport_destination	Transporting_entity
NP.Ext	translocate.v	PP[from].Dep	PP[to].Dep	CNI

Target gene: EG 65192, Rn: Slc27a2; solute carrier family 27 (fatty acid transporter), member 2

(61) the YY1 factor is TRANSLOCATED to the cytoplasm of vaccinia virus infected macrophages CNI

[$_1$ the YY1 factor] is TRANSLOCATED [$_2$ to the cytoplasm of vaccinia virus infected macrophages]

1		2	
Transported_entity	Target LU	Transport_destination	Transporting_entity
NP.Ext	translocate.v	PP[to].Dep	CNI

Target gene: EG 7528, Hs: YY1; YY1 transcription factor

Omission of Transporting_entity also happens via CNI for this LU, as shown below. The following example shows a case of this. The omitted subject of the verb here is actually the GRIF's target gene, 'Arhgef9'. The transported entity is expressed here as the object of the target verb.

(62) TRANSLOCATES gephyrin to submembrane microaggregates CNI

TRANSLOCATES [$_1$ gephyrin] [$_2$ to submembrane microaggregates]

		1	2
Transporting_entity	Target LU	Transported_entity	Transport_destination
CNI	translocate.v	NP.Obj	PP[to].Dep

Target gene: EG 66013, Rn: Arhgef9; Cdc42 guanine nucleotide exchange factor (GEF) 9

There are some cases in the collection in which Transporting_entity is explicitly realized, though not with the LU translocate.v. The following example, using the LU export.v, shows the transporting entity expressed as subject of the target verb and the transported entity as its direct object. Transport locations are also analyzed, the transport origin as a PP and the transport destination as CNI of a case actually mentioned in the GRIF.

(63) Exp5 EXPORTS eEF1A via tRNA from nuclei and synergizes with other transport pathways to confine translation *to the cytoplasm* . CNI

[$_1$ Exp5] EXPORTS [$_2$ eEF1A] via tRNA [$_3$ from nuclei] and synergizes with other transport pathways to confine translation to the cytoplasm .

1		2	3	
Transporting_entity	Target LU	Transported_entity	Transport_origin	Transport_destination
NP.Ext	export.v	NP.Obj	PP[from].Dep	CNI

Target gene: EG 57510, Hs: XPO5; exportin 5

Target gene: EG 13664, Mm: Eif1a; eukaryotic translation initiation factor 1A

Target gene: EG 72322, Mm: Xpo5; exportin 5

Target gene: EG 32970, Dm: official symbol not given; full name not given

Target gene also known as 'Exp5'

Target gene: EG 1964, Hs: EIF1AX; eukaryotic translation initiation factor 1A, X-linked

3.4 Summary

This chapter has provided a detailed presentation of BioFN, an extension of FN to the domain of molecular biology. A definition of the structure of the extension was first offered, and then a specific case study demonstrating the use of BioFN was presented. In the case study, two Frames were defined for different perspectives on the event of intracellular protein transport. A lexicographic sample of two LUs from each Frame demonstrated annotations of GRIFs that make statements about protein transport events. The collection of text annotations chosen for the sample illustrates the range of syntactic and semantic combinatorial possibilities exhibited in language used to discuss events in this domain. In the next chapter, I describe in greater detail how ontological resources of HL can be combined with BioFN for more effective analysis of the data illustrated here. In chapter 5, I show how protein transport analysis and annotations of the sort illustrated here can be used in further analysis of the events described in GRIFs.

Chapter 4

Structure of Resources: Ontological Classes and BioFrameNet Frames

4.1 Introduction

As was reported in chapter 1, the texts covered in this study are taken from a collection of protein transport GRIFs gathered by the Hunter Lab (HL), and analyzed and stored in an ontologically-grounded knowledge base HL has created (Lu et al. 2006; Ogren 2006). This chapter considers differences in structure between the HL knowledge base (HLKB) and that of BioFrameNet. These differences involve higher-level units of analysis, proposed relations between them, and motivations for the relations defined. There are compelling linguistic reasons why for BioFN, the predicates used in the transport GRIFs should be grouped in a different way than they are in HLKB.

The chapter includes the following sections. In section 4.2, I show the five protein transport class definitions and relations proposed in HLKB. In 4.3, I present the predicates that occur in GRIFs associated with these classes, and show GRIF examples in which the same predicate is used in different HLKB classes. Finally, in 4.4 I review the criteria used by FN for grouping LUs in the same Frame and show that they motivate the two protein transport Frames I propose in this study.

4.2 Protein transport classes proposed in HLKB

The structure of HLKB is that of a traditional class-based hierarchically arranged ontology (Noy et al. 2000, Nirenburg and Raskin 2004). Classes in HLKB define and categorize biological phenomena based on explicitly specified features, formally represented as slots for the classes. The HLKB classes examined for this study are ones for defining intracellular protein transport phenomena. Five protein transport classes are defined for this in HLKB. These are listed below in figure (1), with class hierarchy shown with indentation.

(1) HLKB protein transport classes

- *protein transport*
    - *gated nuclear transport*
    - *transmembrane transport*
    - *vesicular transport*
        - *endocytosis*

The top level class defined here, 'protein transport' (PT), includes the following slots:

Slot name	Filler type specification
transport destination	‘cellular component’
transport locations	‘cellular component’
transport origin	‘cellular component’
transport participants	'protein or molecular complex'
transported entity	'protein or molecular complex'
transporting entity	'protein or molecular complex'

TABLE 4.1: HLKB Protein Transport Class Slots, Filler Types

The type specified for fillers of the first three slots is that of cellular component. For fillers of the other slots, the disjunctive type specification 'protein or molecular complex' is provided. Type specifications for a slot’s fillers indicate requirements of class membership for items that are instances of that slot.

The slot definitions for PT, and the type specifications for their fillers, provide the substance of HLKB's formal definition of the phenomenon of protein transport. The human-readable description of this class, taken from the Gene Ontology (GO), is provided below.

(2) HLKB class ‘protein transport’ — Documentation

> The directed movement of a set of molecules and/or molecular complexes into, out of, or within a cell or between cells. (GO:0015031)

Three of the other protein transport classes of HLKB, 'gated nuclear transport', 'transmembrane transport', and 'vesicular transport', are defined as subclasses of PT. Human-readable descriptions of these are listed below.

(3) HLKB subclasses of 'protein transport' class — Documentation

a. 'gated nuclear transport'
The transport of one molecule or molecular complex between the cytosol and nucleus through a nuclear pore.

b. 'transmembrane transport'
The transport of one molecule or molecular complex between the cytosol and a compartment that is topologically distinct from the cytosol and that involves the crossing of a membrane.

c. 'vesicular transport'
The transport of one or more types of molecules and/or molecular complexes from one location to another via a membrane-enclosed transport intermediate. (GO:0016192)

The remaining protein transport class included in HLKB is 'endocytosis', a subclass of the vesicular transport class. This class inherits the slots and type specifications given for vesicular transport, which are inherited from PT. The human-readable description for this subclass is listed below.

(4) HLKB class 'endocytosis' — Documentation

The transport of one or more types of molecules and/or molecular complexes from the extracellular space to an internal cellular compartment via a membrane-enclosed vesicle. (GO:0006897)

Though the descriptions listed above provide information about what makes the phenomena represented by the classes different, the formal structure of the class definitions are otherwise the same. In particular, based on the inheritance relations defined in HLKB, all of the classes include the same slot definitions listed above, with the same type specifications for their fillers.[45]

4.3 Predicates included in HLKB ontology classes

Each GRIF in the collection considered in this study contains at least one transport predicate that semantically heads the description of a protein transport event. Since the transport description is associated with a particular protein transport class, for purposes of presentation, the predicate should also be associated with this class. Below are the forty-four transport predicates observed in the collection, grouped into the five HLKB classes, listed one class at a time. As will be seen in the lists provided here, often a particular predicate can be used for describing more than one kind of transport. A summary table

[45] Inheritance relations like these assert that anything formally specified to be true of the top level class, PT, can be inferred to be true of the classes that inherit from it.

reporting which protein transport class or classes these predicates are used to describe is provided in Appendix 3.

In GRIFs associated with the top level class, 'protein transport', 11 different transport predicates are used. These are shown in figure (5).

(5) Predicates used in GRIFs associated with 'protein transport' class (11)

export.n, mobilization.n, recycle.v, recycling.n, redistribution.n, release.n, relocation.n, target.v, translocate.v, translocation.n, transport.n

The GRIFs associated with the sub-class 'gated nuclear transport' show the greatest number of different transport predicates. The 31 predicates used for describing this kind of transport are shown in figure (6).

(6) Predicates used in GRIFs associated with 'gated nuclear transport' class (31)

delivery.n, distribute.v, efflux.n, enter.v, entry.n, exclude.v, exit.v, export.n, export.v, import.n, import.v, migrate.v, mobilization.n, move.v, movement.n, recruit.v, redistribution.n, relocalize.v, relocate.v, return.v, sequester.v, shift.n, shuttle.v, shuttling.n, target.v, targeting.n, trafficking.n, translocate.v, translocation.n, transport.n, transport.v

In GRIFs associated with the sub-class 'transmembrane-transport', 13 different transport predicates are used. These are shown in figure (7).

(7) Predicates used in GRIFs associated with 'transmembrane transport' class (13)

anchor.v, import.v, move.v, recruitment.n, redistribution.n, release.n, release.v, sequester.v, target.v, targeting.n, translocate.v, translocation.n, transport.n

In GRIFs associated with the subclass 'vesicular transport', 12 different transport predicates are used. These are shown in figure (8).

(8) Predicates used in GRIFs associated with 'vesicular transport' class (12)

exocytosis.n, recycle.v, redistribution.n, relocate.v, sort.v, targeting.n, traffic.n, trafficking.n, translocate.v, translocation.n, transport.n, transport.v

Finally, in GRIFs associated with 'endocytosis', the subclass of 'vesicular transport', 7 different transport predicates are used. These are shown in figure (9).

(9) Predicates used in GRIFs associated with ‘endocytosis’ class (7)

endocytosis.n, internalization.n, internalize.v, target.v, trafficking.n, translocate.v, translocation.n

4.3.1 Same predicate used in describing different kinds of protein transport

As previously noted, many of the predicates listed above are used for describing protein transport events of more than one class. Here we will see two example predicates that show this multi-class possibility. The first example is 'redistribution', a nominal predicate. Figure (10) shows two GRIFs that use this predicate to describe a protein transport type that has been assigned to the top-level protein transport class. Slot fillers realized in the GRIFs are surrounded by square brackets, with the slot name and the name of the class of the filler given in subscript notation. In both these examples we see fillers for the transported entity slot and one or more location slots realized.

(10) GRIFs associated with HLKB *protein transport* class; predicate is ‘redistribution’

a.
biliary anion transport mediated through Mrp2 in the 8-hour cold ischemic liver grafts is via **redistribution** of [Mrp2 $_{\text{TransportedEntity : Protein}}$] from the
[cytoplasm $_{\text{Origin : CellularComponent}}$] to the
[canalicular membrane $_{\text{Destination : CellularComponent}}$]

b.
[estrogen receptor $_{\text{TransportedEntity : Protein}}$] **redistribution** to the
[cytoplasm $_{\text{Destination : CellularComponent}}$] and its interaction with HER2 are important downstream effects of HER2 overexpression

Figure (11) shows uses of the predicate in two GRIFs that describe a type of transport that has been assigned to the 'gated nuclear transport' class.

(11) GRIFs associated with HLKB *gated nuclear transport* class; pred. is ‘redistribution’

a.
Wwox expression triggers **redistribution** of
[[nuclear $_{\text{Origin : CellularComponent}}$] p73 $_{\text{TransportedEntity : Protein}}$] to the
[cytoplasm $_{\text{Destination : CellularComponent}}$] and , hence , suppresses its transcriptional activity .

b.
PC12 cells stably expressing forms of SH2-Bbeta mimicked the ability of NGF to promote **redistribution** of [forkhead $_{\text{TransportedEntity : Protein}}$] (FKHR) to the
[cytoplasm $_{\text{Destination : CellularComponent}}$] .

As with the examples given in (10), slot fillers for transported entity and one or more transport locations are explicitly realized here.

In figures (12) and (13) we see the same transport predicate used in GRIFs that describe protein transport of the classes 'transmembrane transport' and 'vesicular transport', respectively. Though these GRIFs describe protein transport of different classes, we see nonetheless the same types of slot fillers explicitly realized, transported entity and transport locations.

(12) GRIF associated with HLKB *transmembrane transport* class; pred. is ‘redistribution’

> In Parkinson's disease patients , Bcl-xL mRNA expression per dopaminergic neuron is almost double that of controls , an effect that may be mediated by a **redistribution** of [Bcl-xL $_{\text{TransportedEntity : Protein}}$] from the
> [cytosol $_{\text{Origin : CellularComponent}}$] to the
> [outer mitochondrial membrane $_{\text{Destination : CellularComponent}}$] .

(13) GRIF associated with HLKB *vesicular transport* class; pred. is ‘redistribution’

> increase in the concentration of copper in the medium (189 microM) rapidly induces a **redistribution** of the [MNK protein $_{\text{TransportedEntity : Protein}}$] from
> [early sorting endosomes $_{\text{Origin : CellularComponent}}$] , positive for Rab5-myc protein , to
> [late endosomes $_{\text{Destination : CellularComponent}}$] , containing the Rab7-myc protein

In the following figures, we see more examples of GRIFs in which the same predicate is used to describe protein transport of different classes. The predicate in these examples is the verb, 'translocate'. As with the previous examples, fillers for the transported entity slot and different transport locations slots are explicitly realized in these GRIFs.

(14) GRIF associated with HLKB *protein transport* class; pred. is ‘translocate’

> [CERK $_{\text{TransportedEntity : Protein}}$] **translocates** during activation from the
> [cytosol $_{\text{Origin : CellularComponent}}$] to a [lipid raft fraction $_{\text{Destination : CellularComponent}}$]

(15) GRIF associated with HLKB *gated nuclear transport* class; pred. is ‘translocate’

> [IMP1 $_{\text{TransportedEntity : Protein}}$] **translocates** to the [nucleus $_{\text{Destination : CellularComponent}}$] and contains nuclear export signals within the RNA-binding KH2 and KH4 domains .

(16) GRIF associated with HLKB *transmembrane transport* class; pred. is ‘translocate’

> [AIF $_{\text{TransportedEntity : Protein}}$] **translocated** from [mitochondria $_{\text{Origin : CellularComponent}}$] into the [nucleus $_{\text{Destination : CellularComponent}}$] upon nitric oxide exposure . (apoptosis-inducing factor)

(17) GRIF associated with HLKB *vesicular transport* class; pred. is ‘translocate’

in response to Src-dependent activation of phospholipase Cgamma1 , the [Ras guanine nucleotide exchange factor $_{\text{TransportedEntity : Protein}}$] RasGRP1 **translocated** to the [Golgi $_{\text{Destination : CellularComponent}}$] where it activated Ras

(18) GRIF associated with HLKB *endocytosis* class; pred. is ‘translocate’

[cPLA(2)alpha $_{\text{TransportedEntity : Protein}}$] **translocated** to [forming phagosomes $_{\text{Destination : CellularComponent}}$] , surrounding the zymosan particle and completely overlapping with early endosome and plasma membrane markers but only partially overlapping with resident endoplasmic reticulum proteins

In several of the GRIFs shown in the previous figures, we see again cases in which only one of the transport location slots is explicitly realized, further evidence of the CoreSet status these items will have as FEs in BioFN’s protein transport Frames.

4.3.2 Omissions in descriptions of different kinds of protein transport

In addition to similar patterns of slot filler realizations seen in protein transport GRIFs associated with different HLKB protein transport classes, there are similar cases in which FEs are left implicit in the protein transport GRIFs associated with these classes. Figure (19) shows protein transport GRIFs in which the ‘transported entity’ slot filler is left implicit, omissions that will be analyzed as cases of CNI in BioFN. In these cases, the likely interpretation is that the omitted ‘transported entity’ is the GRIF’s target protein.

(19) GRIF associated with HLKB *protein transport* class; pred. is ‘translocation’

activation and **translocation** to [cell membrane $_{\text{Destination : CellularComponent}}$] dependent on protein kinase C

(20) GRIF associated with HLKB *gated nuclear transport* class; pred. is ‘export’

cell density regulates the intracellular localization and function of AhR , because of modulation of [nuclear $_{\text{Origin : CellularComponent}}$] **export** activity

(21) GRIF associated with HLKB *transmembrane transport* class; pred. is ‘translocation’

interactions between Tim44 and mtHsp70 , controlled by polypeptide binding , are required for efficient **translocation** across the [mitochondrial inner membrane $_{\text{TransportLocations : CellularComponent}}$]

(22) GRIF associated with HLKB *vesicular transport* class; pred. is ‘trafficking’

> Sorting nexin 4 and amphiphysin 2 have roles in endocytosis and [intracellular $_{\text{TransportLocations : CellularComponent}}$] **trafficking**

(23) GRIF associated with HLKB *endocytosis* class; pred. is ‘internalize’

> Nociceptin-induced receptor endocytosis mainly occurred via clathrin-coated pits . Mainly **internalized** through the [endosome compartment $_{\text{TransportLocations : CellularComponent}}$] . Receptor phosphorylation necessary for internalization

4.4 FN criteria for grouping LUs in same Frame

As illustrated in the previous section, HL has defined five different classes of protein transport based on specific biological transport mechanisms. For BioFrameNet, I have not created corresponding Frames for each of the classes HL defined, because in general, the language used is the same across all these classes. Having faced this sort of question frequently over the life of the project, FrameNet has in recent releases described explicit criteria to consider when grouping lexical units into Frames. Based on these criteria, it can be seen that there are two kinds of language used for describing protein transport phenomena, both of which are used in the same way across all five HLKB classes. As introduced in Chapter One and described in detail in the previous chapter, I have created two protein transport Frames for BioFrameNet: Protein_transport and Cause_protein_transport. To better understand what is meant by the claim that the "language used is the same" across all HLKB classes, and how I settled on just two protein transport Frames, below is a review of the criteria that FrameNet uses for grouping LUs in Frames[46] and their application to grouping of the transport predicates used in the GRIF collection considered in the study.

General semantics of the Frame

Several of the criteria FN considers for grouping LUs in a Frame involve the general semantics of the scenario evoked by the LUs. One of these is that the basic denotation of the targets in a Frame should be similar, perhaps the most subjective of the criteria considered. When examining the protein transport events classified in HLKB, it is clear that in all cases the notion of directed motion of a protein entity within a cell is indicated, regardless of the particular transport mechanism used. In some cases, though, an agentive participant in the transport event is mentioned. Such cases motivate the creation of a separate causal Frame, one that is related to the basic protein transport Frame.

Another criterion emphasizes that the presuppositions, expectations, and concomitants of the target LUs within a Frame will be shared. This is true of all the predicates denoting the different versions of protein transport described in HLKB classes. Likewise, another

[46] These are listed in the chapter on Frame Development in Ruppenhofer et al. (2006)

criterion notes that in aspectually complex Frames, the targets should all entail the same set of stages and transitions. The example provided in the list of criteria FN considers is that of the different "intended and typical fulfillment stages" seen in 'work on' vs. 'develop', in which the latter is more likely to imply a degree of completion than the former. Here none of the transport predicates included in the GRIF collection systematically indicates differences in this sort of fulfillment stage.

FEs of the Frame

The other criteria FN considers focus more on Frames' FEs. One of these is that all LUs in a Frame must have the same number and types of FEs in both explicit and implicit (NI) contexts. Proposed FEs in BioFN are taken from the slot definitions in the HLKB classes. As noted earlier in this chapter, the top level 'protein transport' class is the source of the slot definitions for all of the classes. In all cases, the key slots are the 'transported entity' slot and the transport location slots, including one or more of 'transport origin', 'transport destination', or 'transport locations'. Because these slots are common across all HLKB protein transport classes, this criterion is met ipso facto in most cases and thus calls for a common Frame. However, the agentive 'transporting entity' slot occurs in only a limited subset of the uses of any given predicate, regardless of protein transport class. These cases belong in a separate causative Frame, linked to the first one via a Frame-to-Frame relation in the Frames' definitions.[47]

The filler types and constraints of the HLKB class slots are all strictly defined. These definitions are shared across all the classes, thus also meeting the criterion of same number and type of FEs for both protein transport Frames proposed.

It is also noted that interrelations between FEs should be the same for all LUs in a Frame. The primary example of this in the Frames proposed here is the CoreSet status of the transport locations, origin and destination, as described in the previous chapter. For all the predicates included in the protein transport GRIF collection, there is variation in which and how many of these FEs are explicitly realized. This is the case for each of the protein transport types classified in HLKB, and thus also for the LUs in both protein transport Frames proposed.

Another criterion considered is that the same FEs will be profiled across all LUs of a Frame. The transported entity FE is usually the profiled item, especially with verbal predicates where it is expressed as the subject or the direct object of the target. For nominal predicates, though, either the transported entity or a transport location, and sometimes even both, can be realized as pre-nominal modifiers. This sort of variation is common across all the predicates, in all kinds of protein transport events described in HLKB classes. The HLKB classes do not display distinctions in profiling (e.g., goods_transfer:buy,sell vs. money_transfer:pay,collect) and perspective (e.g., buy vs. sell), of the sort described in Gawron ’08.

[47] The protein transport class of HLKB also defines a ‘transport participants’ slot, for cases in which GRIFs mention that some entity is involved in a transport event, but its exact role is not certain from the text provided in the GRIF. This slot is used in only 17 transport GRIFs. In most of these cases, it is filled by a constituent outside the local grammatical structure of the target LU, and there is also an explicitly realized filler for the ‘transported entity’ slot.

One other consideration noted is that if there are any pre-specifications given to a Frame's FEs, these should be similar across the LUs of a Frame. The example provided in this case is that of moving entities in the Mass_motion Frame (e.g., 'crowd', 'flock') vs. those in the Self_motion Frame (e.g., 'climb', 'hobble'), where in the former case, the moving entity is a mass theme. This kind of difference in FE pre-specification is not found across the predicates used in transport descriptions across HLKB classes.

In summary, though consideration of the criteria that FN recommends for grouping LUs in Frames does not call for separating predicates used in different HLKB classes into different Frames, creation of a separate causative protein transport Frame is in line with the criteria listed. This separation of a causative protein transport Frame from a basic protein transport Frame spans all the classes of HLKB.

Figures 10 and 11 in chapter 3 present lists of the LUs included in the Protein_transport Frame and the Cause_protein_transport Frame. Example lexical entry reports for 'translocation.n', taken both from Protein_transport and from Cause_protein_transport, are provided in Appendix 4. Each of these reports includes the portion of the LU's summary that is associated with specific HLKB protein transport classes. The summaries in the reports illustrate the general similarity in patterns of FE realization and omission.

4.5 Conclusion

In this chapter I have described the protein transport classes defined in HLKB, and provided the human-readable descriptions offered for each of them. It was shown that while the descriptions offer details of differences in biological properties of the transport mechanisms targeted by the classes, these differences are not otherwise formally represented in the ontological structure of the class definitions. I also listed the predicates used in the protein transport GRIFs associated with each of the HLKB classes, and showed example GRIFs which describe different types of protein transport, and yet use the same predicate and types of slots and fillers. After reviewing the criteria FN uses for grouping LUs in Frames, I argued that the grammar and linguistic semantic properties observed in the protein transport GRIFs are best analyzed with only two protein transport Frames, Protein_transport and Cause_protein_transport. Both of these were defined in chapter 3.

Chapter 5

Conclusions

5.1 Introduction

This study has introduced BioFrameNet (BioFN), an extension of FrameNet to the domain of molecular biology. In this resource, frame semantic analysis is offered for scientific language used in this domain. A key feature of the semantic analysis is its onomasiological perspective, a perspective which focuses on meaning first, and then describes how meanings are encoded in linguistic form. The analysis focuses on the syntactic and semantic combinatorial possibilities characteristic of language used to describe concepts of molecular biology. In doing so, it provides a new perspective on concepts in this domain.

Frame semantic analysis in BioFN is captured in the form of annotations of sentences used for describing events and processes studied in the domain. Annotations provide detailed descriptions of grammatical structures that realize elements of the Frame, and indicators for cases when important frame elements are omitted. The organization and relations of semantic frames proposed in this resource are explicitly specified. This organization covers both relations between the domain-specific Frames within BioFN, as well as relations between BioFN Frames and general language Frames defined in FrameNet.

The remainder of this chapter reviews the case study presented in chapter 3, and summarizes significant findings. I conclude with ideas for future work on BioFN and thoughts on possible linkings with other NLP tools.

5.2 Results

In chapter 3, I presented a case study in which the frame semantics of language used to describe intracellular protein transport phenomena was analyzed. The analysis drew from definitions of this concept asserted in a knowledge base created by the Hunter Lab (HL), of the University of Colorado, in particular, the semantic features proposed for protein transport classes, provided in the form of class ‘slots’. As discussed in chapter 4, though the knowledge base distinguishes five related types of protein transport, the linguistic frame semantics of the predicates and grammatical structures used to express different kinds of transport warrant creation of only two different Frames for BioFN, Protein_transport and Cause_protein_transport.[48] The Protein_transport Frame is related to the general FN Frame 'Motion' by the Frame-to-Frame relation 'Inheritance'

[48] The five intracellular protein transport classes proposed in the HL knowledge base are distinct only in the mechanism of transport, and available there only as non-formal class descriptions. Since mechanism of transport is not an integral portion of linguistic expressions used along with transport predicates, a Frame Element for this has not been proposed in this study.

(Ruppenhofer et al. 2006:104-106). And as suggested by the name, the two Transport Frames are related by the 'Causative_of' relation (ibid.:110). The Frame Elements (FEs) proposed for these Frames are similar to the slot definitions of the protein transport classes of the HL knowledge base (HLKB).

With the definitions of protein Transport Frames in place, I adapted HL's annotations of a large set of protein transport GeneRIFs (GRIFs) for BioFN, putting greater focus on linguistic frame semantics of language used in the texts. From the transport predicates of these GRIFs, I defined forty-four lexical units (LUs), including both nominal predicates and verbal predicates. Often the same predicate shows a causal use in some cases and an inchoative use in others, thus necessitating creation of two different LUs for the same predicate. Annotation reports were created for each of the LUs, providing an organized list of annotated grammatical structures that realize the frame semantics of a protein transport event in GRIFs with the target LU.

The annotation reports produced show analysis of a variety of complex grammatical structures used for expressing protein transport ideas in the GRIFs. As seen in portions of sample annotation reports presented in Chapter 3, typical sources of complexity in scientific writings on molecular biology are multi-word expressions, coordination structures, and inclusion of domain-specific entities in texts. A summary and overview of the various grammatical structures seen in annotation reports is provided in corresponding lexical entry reports, yielding a thorough listing of syntactic and semantic combinatorial possibilities associated with LUs used in the GRIFs. The sets of reports for BioFN Frames and LUs described here are stored and organized on a FN-inspired BioFN website.

An important benefit of the frame semantic approach of BioFN is that it reveals the variety of linguistic structures that are used to realize particular FEs of Transport Frames. Among other things, a thorough analysis of verbal and nominal structures used in protein transport GRIFs revealed that additional information can be gotten from examining the combinatorial possibilities of the two. Though current NLP technologies handle verbs and nouns on their own fairly well, combinations of verbal and nominal structures are not handled as well. BioFN illustrates ample analysis of these sorts of structures, and could thus be used in the development of tools aimed at processing texts with such structures.

An additional benefit provided by analyses done in BioFN is that its annotations allow for clear linking between a lexical resource and domain ontologies, while maintaining a focus on linguistic analysis. Language structures used could suggest when ontology class selections should be modified.

Several challenges came up when creating this domain-specific extension to FN. For syntactic-form pieces, annotating GRIF texts for inclusion in BioFN required special tools to assist with syntactic parsing, e.g., specialized part-of-speech (POS) taggers. For semantic pieces, domain experts were needed to assist in annotating texts. Most of this assistance had already been obtained when HL annotated the GRIFs. In addition, domain resources, e.g., PubMed and publicly available ontologies, offered significant help for these tasks.

5.3 Future work

Useful future work for BioFN includes, first and foremost, creation of new Frames. In addition to their own value, these could provide contextualizing external FEs for Protein Transport Frames. A good candidate for this would be Frames for other related intracellular processes, e.g., protein regulation, gene expression, and metabolism. Also, inclusion of peripheral FEs for any molecular biology Frame would be a good source of contextualizing details specific to this domain.

As has been noted in previous chapters, the GRIFs analyzed in this study usually include information about other processes that are involved in transport events, typically either promoting protein transport or blocking it. The annotations provided in this study can be used in the analysis of these processes, resulting in a fuller analysis of the material included in GRIFs. Pronoun resolution and identification of items omitted via null instantiation can be linked with procedures for filling out, from general knowledge or from material inside the GRIFs, the information that's left implicit. In this way, BioFN analyses can use automatic processes that convert material in the texts into a structured database for the domain.

Dolbey et al. (2006) illustrate an example of using BioFN annotations for automated reasoning. In this work, we used the Protein_transport Frame definition along with the FrameNet Frame, Cause_change_of_position_on_a_scale, to analyze a case of protein transport in which it is asserted that the transport activity of a protein is promoted. A portion of one of the GRIFs shown in Chapter 3 was used for this analysis.[49] The full text of this GRIF is listed again here in figure (1), with the portion of the GRIF that was analyzed indicated with underlining.

(1) SCD1 deficiency specifically increases CTP:choline cytidylyltransferase activity by <u>promoting its translocation into membrane</u> and enhances phosphatidylcholine biosynthesis in liver

The goal of our analysis was to combine Frame semantics of FN and BioFN Frames with links to the domain ontologies of GO, Entrez Gene, and HLKB. These links were expressed in the Description Logic (DL) variant of OWL in order to facilitate inference by means of DL reasoners. With these links in place, an Annotation Ontology that uses the BioFN Ontology as a template was automatically generated. The Annotation Ontology populates the BioFN Ontology with instances of Frames and FEs as well as the actual text data, and satisfies the existential constraints which express Frame and FE relations. Figure (2) shows a part of the Annotation Ontology for the example GRIF.

[49] The particular GRIF used was listed in 3.3.4 as figure (39).

(2) Annotation of portion of GRIF

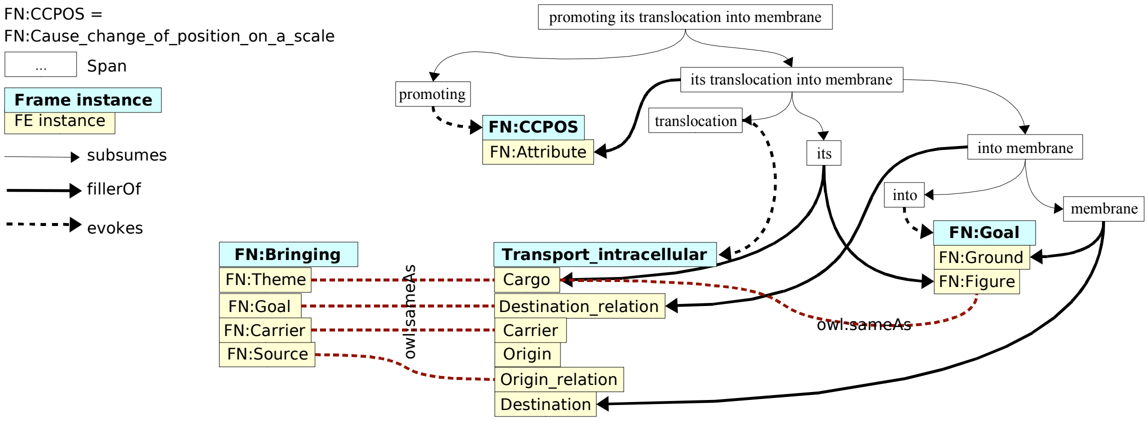
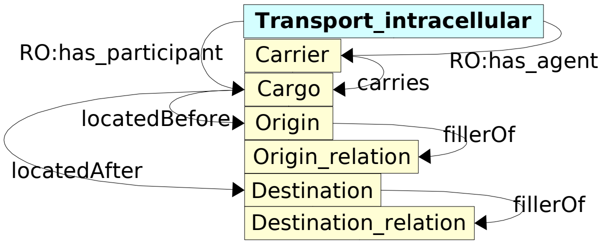


Figure (3) shows additional semantic relations that were generated for the Protein_transport Frame.[50] Several of these were taken from Smith's Relation Ontology (Smith et al. 2005), in order to leverage from their formal definitions.

(3) Semantic relations generated for Protein_transport Frame



With BioFN modeled as an OWL DL ontology populated with BioFN annotations of biomedical texts, natural-language biomedical texts become available for DL-based reasoning.

In addition to the DL-based reasoning described above, BioFN could, ultimately, be used with NLP tools for other important text processing tasks, such as information retrieval, information extraction, and text summarization. An important processing step in many NLP tasks is that of semantic role labeling (SRL), as semantic information is almost always a critical first step for other NLP tasks. Several of the examples shown in the survey of GRIF annotations provided in Chapter 3 demonstrate that FEs are often realized in complex grammatical structures, including coordination and nominal compounds. Results of BioFN efforts thus show that careful linguistic analysis, informed by Frame concepts, could help with SRL.

[50] Some of the names used for BioFN Frames and FEs have been changed since this paper was published.

One way to use BioFN for SRL of domain texts would be to tie BioFN in with other resources and tools that are able to do such processing. Martha Palmer has proposed a strategy for doing this sort of linking, a strategy called SemLink (http://verbs.colorado.edu/semlink/, Loper et al. 2007). Following this strategy, I initiated an effort to link BioFN with PropBank (Kingsbury and Palmer 2002), a resource already used with tools for SRL. Doing this required linking BioFN FEs with PropBank verb arguments. The goal was to use a PropBank-based SRL tool (Loper et al. 2007) for assigning semantic role labels to portions of GRIF texts, and then map the results to BioFN FEs based on the argument linking specified in the previous step.

Several problems came up while attempting this strategy with a domain-specific resource. First, though the SRL tool labeled arguments correctly for some general language verbs, it did not recognize several domain-specific verbs, ones that are very frequently used in transport GRIFs (e.g., 'translocate'). Second, grammatical structures frequently used in domain texts were sometimes not handled successfully. This is likely due to the length limitations imposed on GRIFs, and the frequent use of long compounds this results in. Finally, the particular resource BioFN was linked with, PropBank, does not include nouns as candidate semantic heads for analysis, which resulted in the exclusion of a significant portion of the LUs defined in BioFN.

In the future, these problems could likely be overcome. For example, domain-specific parsers could be used with the SRL tool worked with here. Another solution would be to link with tools that handle processing target nouns and their support verbs.[51] Ultimately, it would be useful to incorporate a lexical resource like BioFN in tools for a variety of important NLP tasks.

[51] Bethard at al. (2008) have implemented just such a SRL tool.

Bibliography

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002. *Molecular Biology of the Cell.* 4th Edition. New York: Garland Science.

Baker, Collin F., Charles J. Fillmore and Beau Cronin. 2003. The Structure of the Framenet Database. *International Journal of Lexicography*, Volume 16.3: 281-296.

Bentivogli, Luisa, Andrea Bocco, and Emanuele Pianta. 2004. ArchiWordNet: Integrating WordNet with Domain-Specific Knowledge. *Proceedings of the Second Global WordNet Conference*: 39-46. Brno, Czech Republic.

Bethard, Steven, Lu Zhiyong, James H. Martin, and Lawrence Hunter. 2008. Semantic Role Labeling for Protein Transport Predicates. *BMC Bioinformatics*, Jun 11;9(1):277.

BioFN website: http://dolbey.us/MediaWiki/index.php?title=BioFrameNet

Buitelaar, Paul and Sacaleanu Bogdan. 2002. Extending Synsets with Medical Terms. *Proceedings of the 1st International WordNet Conference*. Mysore, India.

Center for Computational Pharmacology (CCP) website: http://compbio.uchsc.edu/Hunter_lab/CCP_website/

Cohen, K. Bretonnel, Andrew E. Dolbey, George K. Acquaah-Mensah, and Lawrence Hunter. 2002. Contrast and variability in gene names. *Proceedings of the workshop on natural language processing in the biomedical domain,* 14-20. Association for Computational Linguistics.

Cohen, K. Bretonnel and Lawrence Hunter. 2005. Natural Language Processing and Systems Biology. *Artificial Intelligence Methods and Tools for Systems Biology*: 147-174. Werner Dubitzky and Francisco e Azuaj (Eds.) Dordrecht, the Netherlands; Norwell, MA: Springer.

Cohen, K. Bretonnel and Lawrence Hunter. 2006. A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics*, Vol. 7, No. Suppl 3.

Dolbey, Andrew, Michael Ellsworth, and Jan Scheffczyk. 2006. BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. *KR-MED, Biomedical Ontology in Action*.

Dorr. 2001. Database available at: http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html

Ellsworth, Michael, Katrin Erk, Paul Kingsbury, and Sebastian Padó. 2004. PropBank, SALSA, and FrameNet: How design determines product. *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*. Lisbon: European Language Resources Association.

Entrez Gene website: http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

Fellbaum, Christiane (Ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum, Christiane, Udo Hahn, and Barry Smith. 2006. Towards new information resources for public health. WordNet to MedicalWordNet. *Journal of Biomedical Informatics* 39(3): 321-332. Elsevier.

Fillmore, Charles J. 1982. Frame semantics. *Linguistics in the Morning Calm:* 111-137. Seoul: Hanshin Publishing Co.

Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, Vol. 6.2: 222-254.

Fillmore, Charles J. and Beryl T. S. Atkins. 1992. Towards a frame-based organization of the lexicon: The semantics of RISK and its neighbors. Lehrer, A and E. Kittay (Eds.) *Frames, Fields, and Contrast: New Essays in Semantics and Lexical Organization.* Hillsdale: Lawrence Erlbaum Associates, 75-102.

Fillmore, Charles J., Christopher R. Johnson and Miriam R.L. Petruck. 2003. Background to Framenet. *International Journal of Lexicography*, Vol 16.3: 235-250.

Fillmore, Charles J., Miriam R.L. Petruck, Josef Ruppenhofer, and Abby Wright. 2003. Framenet in Action: The Case of Attaching. *International Journal of Lexicography*, Vol 16.3: 297-332.

FrameNet website: http://framenet.icsi.berkeley.edu/

Friedman, Carol, Pauline Kra and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics* 35: 222-235.

Gawron, Jean Mark. 2008. Frame Semantics. (Unpublished, but available at: http://www.hf.uib.no/forskerskole/new_frames_intro.pdf)

Geeraerts, Dirk. 2006. Towards a multifactorial grammar. Bergen, Norway. University of Leuven. RU Quantitative Lexicology and Variational Linguistics.

Gene Ontology (GO) website: http://www.geneontology.org/

Gene References into Function (GRIF) website: http://www.ncbi.nlm.nih.gov/projects/GeneRIF/

Grondelaers, Stefan and Dirk Geeraerts. 2003. Towards a pragmatic model of cognitive onomasiology. *Cognitive Approaches to Lexical Semantics*: 67-92. Hubert Cuyckens, RenÃ Dirven, and John Taylor (Eds.) Berlin/New York: Mouton de Gruyter.

Harris, Zellig. 1982. *A grammar of english on mathematical principles*. New York: Wiley.

Harris, Zellig. 1991. *A theory of language and information: a mathematical approach*. Oxford: Clarendon Press.

Kictionary website: http://www.kicktionary.de/index.html

Kingsbury, Paul and Martha Palmer. 2002. From TreeBank to PropBank. *Third International Conference on Language Resources and Evaluation, LREC-02*. Las Palmas, Canary Islands, Spain: European Language Resources Association.

Kipper Schuler, Karin. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania.

Koch, Peter. 2008. Cognitive onomasiology and lexical change: Around the eye. *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations*: 107-137. Martine Vanhove (Ed.) Amsterdam: John Benjamins Publishing Company.

Lodish, Harvey, Arnold Berk, Chris A. Kaiser, Monty Krieger, Matthew P. Scott, Anthony Bretscher, Hidde Ploegh, and Paul Matsudaira. 2007. Molecular Cell Biology; 6th edition. W. H. Freeman.

Loper, Edward, Szu-Ting Yi, and Martha Palmer 2007. Combining Lexical Resources: Mapping between PropBank and VerbNet. *Proceedings of the 7th International Workshop on Computational Linguistics,* Tilburg, the Netherlands.

Lu, Zhiyong, K. Bretonnel Cohen, and L. Hunter. 2006. Finding GeneRIFs via gene ontology annotations. *Proceedings of Pacific Symposium on Biocomputing.*

Marcus, Mitch. 1994. The Penn treebank: A revised corpus design for extracting predicate-argument structure. *Proceedings of the ARPA Human Language Technology Workshop*. Princeton, NJ.

Mel’čuk , Igor. 1998. Collocations and lexical functions. In *Phraseology: Theory, Analysis, and Applications*: 23-54. A. P. Cowie (Ed.) Oxford Studies in Lexicography and Lexicology. Oxford: Clarendon Press.

Miller, George A. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38 (11): 39 - 41.

Nakov, Preslav and Marti Hearst. 2006. Using verbs to characterize noun-noun relations. *Proceedings of the Twelfth International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA),* Bulgaria.

National Center for Biotechnology Information (NCBI) website: http://www.ncbi.nlm.nih.gov/projects/GeneRIF/

National Institutes of Health (NIH) website: http://www.nih.gov/

National Library of Medicine (NLM) website: http://www.nlm.nih.gov/

Nirenburg, Sergei and Victor Raskin. 2004. *Ontological Semantics*. Cambridge, MA: MIT Press.

Noy, Natalya F., Monica Crubézy, Ray W. Fergerson, Holger Knublauch, Samson W. Tu, Jennifer Vendetti, and Mark A. Musen. 2003. Protégé-2000: An Open-Source Ontology-Development and Knowledge-Acquisition Environment: AMIA Open Source Expo.

Ogren, Philip V. 2006. Knowtator: A Protégé plug-in for annotated corpus construction. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume*: 273-275. New York City.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, Vol. 31, No. 1, 71-106. Cambridge, MA: MIT Press.

PASBio website: http://research.nii.ac.jp/~collier/projects/PASBio/

Petruck, Miriam R. L. 1996. Frame Semantics. Jef Verschueren, Jan-Ola Östman, Jan Blommaert, and Chris Bulcaen (Eds.). *Handbook of Pragmatics.* Philadelphia: John Benjamins.

Pradhan, Sameer, Honglin Sun, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2004. Parsing Arguments of Nominalizations in English and Chinese. *Proceedings of NAACL-HLT*. Boston, Mass.

PubMed website: http://www.pubmed.gov/

Rosario, Barbara and Marti A. Hearst. 2004. Classifying semantic relations in bioscience texts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

Rosario, Barbara and Marti A. Hearst. 2005. Multi-way Relation Classification: Application to Protein-Protein Interactions. *Proceedings of 2005 Conference on Empirical Methods in Natural Language Processing*. Vancouver, B.C., Canada (EMNLP 2005).

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice.* Website: http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126

Schmidt, Thomas. (2008). The Kicktionary Revisited. Angelika Storrer, Alexander Geyken, Alexander Siebert, and Kay-Michael Würzner (Eds.) *Text Resources and Lexical Knowledge:* 239-252. Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008. Berlin: Mouton de Gruyter.

SemLink website: http://verbs.colorado.edu/semlink/

Smith, Barry and Christiane Fellbaum. 2004. Medical WordNet: a new methodology for the construction and validation of information resources for consumer health. *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*.

Smith, Barry, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L. Rector, and Cornelius Rosse. 2005. Relations in biomedical ontologies. *Genome Biology.* 6(5):R46.1-R46.15.

Wattarujeekrit, Tuangthong. 2005. Exploring Semantic Roles for Named Entity Recognition in the Molecular Biology Domain. PhD thesis. Department of Informatics, School of Multidisciplinary Sciences. The Graduate University for Advanced Studies.

Wattarujeekrit, Tuangthong and Nigel Collier. 2005. Exploring Predicate-Argument Relations for Named Entity Recognition in the Molecular Biology Domain. To appear in *Proceedings of the Eighth International Conference on Discovery Science* (DS'05), Marina Mandarin Hotel, Singapore.

Wattarujeekrit, Tuangthong, Parantu K. Shah, and Nigel Collier. 2004. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, Vol. 5, 155.

Wermter, Joachim and Udo Hahn. 2004. Really, Is Medical Sublanguage That Different? Experimental Counter-evidence from Tagging Medical and Newspaper Corpora. M. Fieschi et al. (Eds). *MEDINFO 2004.* Amsterdam: IOS Press. IMIA.

Yi, Szu-ting and Martha Palmer. 2004. Pushing the boundaries of semantic role labeling with svm. *Proceedings of the International Conference on Natural Language Processing*.

Zholkovsky, Alexander. 1984. *Themes and Texts. Toward a Poetics of Expressiveness*. Foreword by Jonathan Culler. Ithaca and London: Cornell University Press.

Appendix 1: Data from Protein_transport Frame

Annotation highlights of two lexical units presented: translocation.n, translocate.v

Verbal lexical unit: Protein_transport.translocate.v

GRIF Annotations (*seen in 5 Valence Patterns*)

Valence Pattern

Seen in 2 of 10 GRIFs with this LU (20.0%)

Transport_destination	Transport_origin	Transported_entity
PP[to]	PP[from]	NP
Dep	Dep	Ext

GRIF 105980

EntrezGene ID: 21804 [symbol: Tgfb1i1]

PubMed Doc ID: 1571374

hic-5 is a mediator of tensional force , TRANSLOCATING directly from focal adhesions to actin stress fibers upon mechanical stress and regulating the contractile capability of cells in the stress fibers

GRIF 93018

EntrezGene ID: 64781 [symbol: CERK]

PubMed Doc ID: 1589989

CERK TRANSLOCATES during activation from the cytosol to a lipid raft fraction

Valence Pattern

Seen in 1 of 10 GRIFs with this LU (10.0%)

Transport_destination	Transport_origin	Transported_entity
PP[to]	CNI	NP
Dep	--	Ext

GRIF 32162

EntrezGene IDs: 6915 [symbol: TBXA2R] , 4831 [symbol: NME2]

PubMed Doc ID: 1497620

Nm23-H2 had *a cytoplasmic and nuclear localization* but was induced to TRANSLOCATE to the plasma membrane upon stimulation of thromboxane A2 receptor beta to show extensive co-localization with the receptor . CNI

Valence Pattern

`Seen in 5 of 10 GRIFs with this LU (50.0%)`

Transport_destination	Transported_entity
PP[to]	NP
Dep	Ext

GRIF 38356

EntrezGene ID: 243912 [symbol: Hspb6]

PubMed Doc ID: 1510529

Recombinant mouse Hsp20 TRANSLOCATES to and interacts with actin cytoskeleton in response to isoproterenol stimulation and prevents ss-agonist-induced apoptosis in adult rat cardiomyocytes .

GRIF 56775

EntrezGene ID: 351 [symbol: APP]

PubMed Doc ID: 1265984

after A beta binds to raft-like membranes composed of monosialoganglioside GM1/cholesterol/sphingomyelin (1/1/1) , the protein can TRANSLOCATE to the phosphatidylcholine membranes to which soluble A beta does not bind

GRIF 107345

EntrezGene IDs: 117086 [symbol: Stim1] , 32556 [symbol: Stim]

PubMed Doc ID: 1620837

proposal that STIM1 functions as the missing link between Ca2+ store depletion and store-operated calcium influx , serving as a Ca2+ sensor that TRANSLOCATES upon store depletion to the plasma membrane to activate CRAC channels

GRIF 96478

EntrezGene ID: 18708 [symbol: Pik3r1]

PubMed Doc ID: 1604351

In response to IGF-1 , but not to PDGF signaling , EGFP-p85alpha TRANSLOCATES to discrete foci in the cell

GRIF 54622

EntrezGene ID: 60626 [symbol: RIC8A]

PubMed Doc ID: 1265264

We identified a protein member of the synembryn family as one of the interacting proteins in human brain . Gqalpha also interacts with synembryn . Synembryn TRANSLOCATES to the plasma membrane in response to carbachol and isoproterenol .

Valence Pattern

Seen in 1 of 10 GRIFs with this LU (10.0%)

Transport_locations	Transported_entity
PP[across]	NP
Dep	Ext

GRIF 4417

EntrezGene ID: 1991 [symbol: ELA2]

PubMed Doc ID: 1470596

An important part of the antimicrobial mechanism of neutrophil elastase may be a periplasmic bacteriostatic effect of protease that has TRANSLOCATED across the damaged outer membrane .

Valence Pattern Seen in 1 of 10 GRIFs with this LU (10.0%)	
Transport_locations	Transported_entity
CNI --	CNI --

GRIF 6978 (10 of 10)

EntrezGene ID: 2185 [symbol: PTK2B]

PubMed Doc ID: 1207725

By rapidly TRANSLOCATING to the vicinity of the immune synapse after T cell receptor stimulation , Pyk2 plays an essential role in T cell activation and polarized secretion of cytokines . CNI CNI

Nominal lexical unit: Protein_transport.translocation.n

GRIF Annotations (*seen in 26 Valence Patterns*)

Valence Pattern

Seen in 3 of 207 GRIFs with this LU (1.45%)

Transport_destination	Transport_origin	Transported_entity
PP[to]	PP[from]	PP[of]
Dep	Dep	Dep

GRIF 12560

EntrezGene ID: 5170 [symbol: PDPK1]

PubMed Doc ID: 1214768

Data show that the TRANSLOCATION of 3-phosphoinositide-dependent protein kinase-1 from cytosol to the plasma membrane is critical for Akt and glycogen synthase kinase-3 activation .

GRIF 87171

EntrezGene ID: 5588 [symbol: PRKCQ]

PubMed Doc ID: 1574985

increased TRANSLOCATION from the cytoplasm to the membrane of protein kinase theta , a T cell signaling molecule that colocalizes with the TCR within the supramolecular activation cluster

Valence Pattern

Seen in 2 of 207 GRIFs with this LU (0.97%)

Transport_destination	Transport_origin	Transported_entity
PP[to]	PP[from]	N
Dep	Dep	Dep

GRIF 95720

EntrezGene ID: 1991 [symbol: ELA2]

PubMed Doc ID: 1614814

HNE causes protein kinase C (PKC) activation and TRANSLOCATION from cytosol to plasma membrane , required for HNE-induced ROS generation and other responses

GRIF 23061

EntrezGene ID: 84021 [symbol: Irs3]

PubMed Doc ID: 1285028

IRS-3 expression blocked glucose/IGF-1 induced IRS-2 TRANSLOCATION from the cytosol to the plasma membrane , dampening IRS-2/IGF-1R interaction and subsequent activation of the PI3K/PKB/GSK3 signaling pathway

---

Valence Pattern

Seen in 1 of 207 GRIFs with this LU (0.48%)

Transport_destination	Transport_origin	Transported_entity
PP[to]	PP[of]	PP[of]
Dep	Dep	Dep

GRIF 22630

EntrezGene ID: 20525 [symbol: Slc2a1]

PubMed Doc ID: 1286957

IL-3 caused TRANSLOCATION of intracellular GLUT1 transporters to the cell surface

---

Valence Pattern

Seen in 1 of 207 GRIFs with this LU (0.48%)

Transport_destination	Transport_origin	Transported_entity
PP[to]	CNI	Poss
Dep	--	Gen

GRIF 44872

EntrezGene ID: 7133 [symbol: TNFRSF1B]

PubMed Doc ID: 1278660

TNFR2 activates *cytosolic* phospholipase A2 (cPLA2) by causing its TRANSLOCATION to plasma membrane and perinuclear subcellular regions and by causing an increase in intracellular calcium that may contribute to the translocation and activation of cPLA2 . CNI

Valence Pattern

Seen in 7 of 208 GRIFs with this LU (3.37%)

Transport_destination	Transported_entity
N	PP[of]
Dep	Dep

GRIF 73393

EntrezGene ID: 5728 [symbol: PTEN]

PubMed Doc ID: 1280814

phosphorylation/dephosphorylation of the C-terminal region of PTEN serves as an electrostatic switch that controls the membrane TRANSLOCATION of the protein

GRIF 41238

EntrezGene ID: 3937 [symbol: LCP2]

PubMed Doc ID: 1249642

SLP-76 is essential for NF-kappa B activation and lipid raft TRANSLOCATION of protein kinase C theta and the I kappa B kinase complex .

Valence Pattern

Seen in 2 of 207 GRIFs with this LU (0.97%)

Transport_destination	Transported_entity
PP[to]	NP
Dep	Ext

GRIF 51106

EntrezGene ID: 20528 [symbol: Slc2a4]

PubMed Doc ID: 1531416

These results suggest that GLUT4 requires TRANSLOCATION to the plasma membrane , as well as activation at the plasma membrane , to initiate glucose uptake , and both of these steps normally require PI 3-kinase activation .

GRIF 41813 (70 of 207)

EntrezGene ID: 117268 [symbol: Khdrbs1]

PubMed Doc ID: 1499693

Sam68 undergoes activity-responsive TRANSLOCATION to the soma and dendrites of hippocampal neurons in primary culture

Valence Pattern

Seen in 15 of 207 GRIFs with this LU (7.25%)

Transport_destination	Transported_entity
PP[to]	N
Dep	Dep

GRIF 78741

EntrezGene IDs: 18762 [symbol: Prkcz] , 5590 [symbol: PRKCZ]

PubMed Doc ID: 1563045

stromal cell-derived factor 1 triggered PKC-zeta phosphorylation , TRANSLOCATION to the plasma membrane , and kinase activity

GRIF 68210

EntrezGene IDs: 5595 [symbol: MAPK3] , 5594 [symbol: MAPK1]

PubMed Doc ID: 1506908

insulin stimulates Na(+),K(+)-ATPase activity and TRANSLOCATION to plasma membrane in HSMCs via phosphorylation of the alpha-subunits by ERK1/2 mitogen-activated protein kinase .

GRIF 1590

EntrezGene ID: 13139 [symbol: Dgka]

PubMed Doc ID: 1262653

A role for DGK alpha in T cell activation is confirmed by the rapid DGK alpha TRANSLOCATION to the membrane fraction , together with the increase in enzyme activity that follows T cell activation in vivo .

Valence Pattern

```
Seen in 3 of 207 GRIFs with this LU (1.45%)
```

Transport_destination	Transported_entity
PP[into] Dep	PP[of] Dep

GRIF 31868

EntrezGene ID: 59323 [symbol: Erbb4]

PubMed Doc ID: 1271692

Neuregulin stimulation causes TRANSLOCATION of ErbB4 into lipid rafts and these are necessary for signaling by ErbB4

Valence Pattern

```
Seen in 1 of 207 GRIFs with this LU (0.48%)
```

Transport_destination	Transported_entity
PP[into] Dep	Poss Gen

GRIF 100083

EntrezGene IDs: 20249 [symbol: Scd1] , 13026 [symbol: Pcyt1a]

PubMed Doc ID: 1582948

SCD1 deficiency specifically increases CTP:choline cytidylyltransferase activity by promoting its TRANSLOCATION into membrane and enhances phosphatidylcholine biosynthesis in liver

Valence Pattern

Seen in 3 of 207 GRIFs with this LU (1.45%)

Transport_destination	Transported_entity
PP[to]	Poss
Dep	Gen

GRIF 106329

EntrezGene IDs: 25246 [symbol: Bsg] , 25027 [symbol: Slc16a1], 9122 [symbol: SLC16A4]

PubMed Doc ID: 1591724

inhibition of MCT1 and MCT4 activity by pCMBS is mediated through its binding to CD147 , acting as an ancillary protein required to maintain the catalytic activity of MCTs 1 and 4 , as well as for their TRANSLOCATION to the plasma membrane

GRIF 95754 (36 of 207)

EntrezGene ID: 15507 [symbol: Hspb1]

PubMed Doc ID: 1573110

HSP25 binds to protein kinase C delta to inhibit its kinase activity and TRANSLOCATION to the membrane , which results in reduced cell death .

Valence Pattern

Seen in 2 of 207 GRIFs with this LU (0.97%)

Transport_destination	Transported_entity
PP[into]	N
Dep	Dep

GRIF 44185

EntrezGene ID: 355 [symbol: FAS]

PubMed Doc ID: 1474544

TCR restimulation of activated CD4(+) T cells resulted in Fas TRANSLOCATION into lipid raft microdomains before binding FasL , rendering these cells sensitive to apoptosis

Valence Pattern

Seen in 24 of 207 GRIFs with this LU (11.59%)

Transport_destination	Transported_entity
PP[to]	PP[of]
Dep	Dep

GRIF 96442

EntrezGene ID: 360820 [symbol: Pxn]

PubMed Doc ID: 1591174

In mesenteric arteries an intact cytoskeleton and force development are necessary for the TRANSLOCATION of PYK2 and paxillin to the membrane

GRIF 19659

EntrezGene IDs: 29172 [symbol: Aqp8] , 24952 [symbol: Gcg]

PubMed Doc ID: 1277402

Glucagon induces protein kinase A and microtubule-dependent TRANSLOCATION of AQP8 water channels to hepatocyte canalicular plasma membrane , leading to increase in membrane water permeability .

Valence Pattern

Seen in 1 of 207 GRIFs with this LU (0.48%)

Transport_destination	Transported_entity
PP[onto]	PP[of]
Dep	Dep

GRIF 43352

EntrezGene ID: 3084 [symbol: NRG1]

PubMed Doc ID: 1515941

Stimulation of cells with neuregulin-1beta induced Ser-978 dephosphorylation , TRANSLOCATION of SSH1L onto F-actin-rich lamellipodia , and cofilin dephosphorylation .

Valence Pattern

Seen in 1 of 207 GRIFs with this LU (0.48%)

Transport_destination	Transported_entity
N	N
Dep	Dep

GRIF 46600

EntrezGene ID: 54245 [symbol: Crk]

PubMed Doc ID: 1219815

The Y221 site in transfected rat CrkII regulates Rac membrane TRANSLOCATION upon cell adhesion , which is necessary for activation of downstream Rac signaling pathways .

Valence Pattern

Seen in 2 of 207 GRIFs with this LU (0.97%)

Transport_destination	Transported_entity
A	PP[of]
Dep	Dep

GRIF 100939

EntrezGene IDs: 280943 [symbol: TNF] , 282527 [symbol: TNFRSF1A], 504707 [symbol: TRADD]

PubMed Doc ID: 1531075

TNF binding induces release of AIP1 (DAB2IP) from TNFR1 , resulting in cytoplasmic TRANSLOCATION and concomitant formation of an intracellular signaling complex comprised of TRADD , RIP1 , TRAF2 , and AIP1 .

Valence Pattern

Seen in 1 of 207 GRIFs with this LU (0.48%)

Transport_destination	Transported_entity
A	NP
Dep	Ext

GRIF 107432

EntrezGene ID: 19674 [symbol: Rcvrn]

PubMed Doc ID: 1596139

recoverin undergoes light-dependent intracellular TRANSLOCATION in mouse rod photoreceptors

---

Valence Pattern

Seen in 1 of 207 GRIFs with this LU (0.48%)

Transport_destination	Transported_entity
A	CNI
Dep	--

GRIF 75741

EntrezGene ID: 6915 [symbol: TBXA2R]

PubMed Doc ID: 1458363

results indicate that oxidative stress induces maturation and stabilization of *the thromboxane A(2)Receptor beta protein* probably by intracellular TRANSLOCATION CNI

Valence Pattern

Seen in 5 of 207 GRIFs with this LU (2.42%)

Transport_destination	Transported_entity
PP[to]	CNI
Dep	--

GRIF 95847

EntrezGene ID: 319757 [symbol: Smo]

PubMed Doc ID: 1613607

Hh-dependent TRANSLOCATION to cilia is essential for *Smo* activity , suggesting that *Smo* acts at the primary cilium CNI

GRIF 56358 (74 of 207)

EntrezGene ID: 8877 [symbol: SPHK1]

PubMed Doc ID: 1212438

activation and TRANSLOCATION to cell membrane dependent on protein kinase C CNI

GRIF 82183 (75 of 207)

EntrezGene ID: 25660 [symbol: Cd28]

PubMed Doc ID: 1528053

TRANSLOCATION to lipid rafts may play an important role in CD28 signaling . CNI

Valence Pattern

Seen in 1 of 207 GRIFs with this LU (0.48%)

Transport_locations	Transported_entity
PP[through]	N
Dep	Dep

GRIF 9428

EntrezGene IDs: 84592 [symbol: Cetn1] , 26369 [symbol: Cetn1]

PubMed Doc ID: 1534765

Cen1 and Cen2 contribute to the centrin-transducin complex and potentially participate in the regulation of transducin TRANSLOCATION through the photoreceptor cilium

---

Valence Pattern

Seen in 2 of 207 GRIFs with this LU (0.97%)

Transport_locations	Transported_entity
PP[across]	PP[of]
Dep	Dep

GRIF 12988

EntrezGene ID: 5830 [symbol: PEX5]

PubMed Doc ID: 1241143

Data suggest that TRANSLOCATION of PTS1-containing proteins across the peroxisomal membrane occurs concomitantly with formation of the Pex5p-Pex14p membrane complex and that this is probably the site from which Pex5p leaves the peroxisomal compartment .

Valence Pattern

Seen in 1 of 207 GRIFs with this LU (0.48%)

Transport_locations	Transported_entity
PP[in]	N
Dep	Dep

GRIF 22175

EntrezGene IDs: 16000 [symbol: Igf1] , 12389 [symbol: Cav1]

PubMed Doc ID: 1213560

IGF-I induces caveolin 1 tyrosine phosphorylation and TRANSLOCATION in the lipid rafts .

Valence Pattern

Seen in 2 of 207 GRIFs with this LU (0.97%)

Transport_locations	Transported_entity
PP[at]	PP[of]
Dep	Dep

GRIF 103604

EntrezGene IDs: 14083 [symbol: Ptk2] , 5747 [symbol: PTK2]

PubMed Doc ID: 1603960

We propose a model in which removal of FERM-mediated auto-inhibition is important to increase FAK catalytic activity but the TRANSLOCATION and clustering of this enzyme at the focal adhesions is required for maximal phosphorylation at Tyr-397 .

Valence Pattern

Seen in 34 of 207 GRIFs with this LU (16.43%)

Transport_locations	Transported_entity
CNI	PP[of]
--	Dep

GRIF 82179

EntrezGene ID: 56616 [symbol: DIABLO]

PubMed Doc ID: 1236434

TRANSLOCATION of Smac along with cytochrome c and other mitochondrial pro-apoptotic proteins represent important regulatory checkpoints for *mitochondria*-mediated apoptosis CNI

GRIF 30670 (83 of 207)

EntrezGene IDs: 16367 [symbol: Irs1] , 16001 [symbol: Igf1r], 384783 [symbol: Irs2]

PubMed Doc ID: 1255475

Mutations in the tyrosine kinase domain of the IGF-IR abrogate TRANSLOCATION of the IRS-2 and IRS-2 proteins . CNI

GRIF 13624 (90 of 207)

EntrezGene IDs: 207 [symbol: AKT1] , 24185 [symbol: Akt1]

PubMed Doc ID: 1211202

Different cellular localization , TRANSLOCATION , and insulin-induced phosphorylation of PKBalpha in HepG2 cells and hepatocytes CNI

GRIF 81506 (107 of 207)
EntrezGene ID: 27101 [symbol: CACYBP]
PubMed Doc ID: 1289529

the TRANSLOCATION of CacyBP during the retinoic acid-induced differentiation of neuroblastoma SH-SY5Y cells suggested that this protein might play a role in neuronal differentiation CNI

Valence Pattern

```
Seen in 5 of 207 GRIFs with this LU (2.42%)
```

Transport_locations	Transported_entity
CNI	Poss
--	Gen

GRIF 34608 (116 of 207)
EntrezGene ID: 5579 [symbol: PRKCB1]
PubMed Doc ID: 1205690

PKC beta II expressed in human neutrophils can phosphorylate p47phox and induce both its TRANSLOCATION and NADPH oxidase activation as well as the binding of p47phox to the cytosolic fragment of p22phox . CNI

GRIF 43858 (120 of 207)
EntrezGene IDs: 5580 [symbol: PRKCD] , 5581 [symbol: PRKCE]
PubMed Doc ID: 1187742

Switch chimeras , containing the C1B from epsilonPKC in the context of deltaPKC (delta(epsilonC1B)) and vice versa (epsilon(deltaC1B)) , were generated and tested for their TRANSLOCATION in response to ceramide and arachidonic acid . CNI

Valence Pattern

Seen in 83 of 207 GRIFs with this LU (40.10%)

Transport_locations	Transported_entity
CNI	N
--	Dep

GRIF 3747 (123 of 207)

EntrezGene IDs: 7430 [symbol: VIL2] , 6550 [symbol: SLC9A3]

PubMed Doc ID: 1553158

Akt2-dependent ezrin phosphorylation leads to NHE3 TRANSLOCATION and activation CNI

GRIF 88840 (125 of 207)

EntrezGene ID: 363855 [symbol: Map2k7]

PubMed Doc ID: 1581685

Map2k7 activation , TRANSLOCATION and binding to Jun N-terminal kinase (JNK)-interacting protein (JIP)-1 is closely associated with reactive oxygen species and might play a pivotal role in the activation of JNK signaling in brain ischemic injury . CNI

GRIF 107414 (128 of 207)

EntrezGene IDs: 6609 [symbol: SMPD1] , 5599 [symbol: MAPK8]

PubMed Doc ID: 1576973

raft-associated acid sphingomyelinase and JNK activation and TRANSLOCATION are induced by UV-C light on a nuclear signal CNI

GRIF 88677 (135 of 207)

EntrezGene ID: 25551 [symbol: Slc2a3]

PubMed Doc ID: 1532087

Inhibition of mitochondrial respiration by NO triggered a rapid , cGMP-independent enhancement of GLUT3-mediated glucose uptake through a mechanism that did not involve transporter TRANSLOCATION CNI

GRIF 40834 (146 of 207)

EntrezGene ID: 6284 [symbol: S100A13]

PubMed Doc ID: 1503349

S100A13 protein TRANSLOCATION in response to extracellular S100 is mediated by receptor for advanced glycation endproducts in human endothelial cells CNI

GRIF 41362 (151 of 207)
EntrezGene ID: 89829 [symbol: Socs3]
PubMed Doc ID: 1551408

SOCS-3 participates , as a late event , in the negative cross-talk between angiotensin II and insulin , producing an inhibitory effect on insulin-induced glucose transporter-4 TRANSLOCATION . CNI

GRIF 10405 (160 of 207)
EntrezGene ID: 78975 [symbol: Prkaa2]
PubMed Doc ID: 1282962

Contraction-induced fatty acid translocase/CD36 TRANSLOCATION in rat cardiac myocytes is mediated through this enzyme's signaling . CNI

GRIF 78173 (186 of 207)
EntrezGene ID: 81649 [symbol: Mapk14]
PubMed Doc ID: 1137488

signal pathway mediating 5-lipoxygenase TRANSLOCATION and cell death CNI

Valence Pattern

`Seen in 4 of 207 GRIFs with this LU (1.93%)`

Transport_locations	Transported_entity
CNI	CNI
--	--

GRIF 61278 (205 of 207)
EntrezGene ID: 79212 [symbol: Slc6a1]
PubMed Doc ID: 1276415

data suggest that the extracellular end of transmembrane domain 7 of *GABA transporter 1* not only undergoes conformational changes critical for the TRANSLOCATION process but also plays a role in regulating the conformational equilibrium CNI CNI

GRIF 102727 (207 of 207)
EntrezGene ID: 948353 [symbol: dsbA]
PubMed Doc ID: 1593716

These results suggest that *DsbA* uses not only the signal recognition particle targeting pathway but also a special route of TRANSLOCATION through the translocon . CNI CNI

Appendix 2: Data from Cause_protein_transport Frame

Annotation highlights of two lexical units presented: translocation.n, translocate.v

Verbal lexical unit: Cause_protein_transport.translocate.v

GRIF Annotations (*seen in 6 Valence Patterns*)

Valence Pattern

Seen in 1 of 17 GRIFs with this LU (5.88%)

Transport_origin	Transport_destination	Transported_entity	Transporting_entity
PP[from]	PP[to]	NP	CNI
Dep	Dep	Ext	--

GRIF 24403

EntrezGene ID: 65192 [symbol: Slc27a2]

PubMed Doc ID: 1200989

In response to an increase of cellular cholesterol , fatty acid transporter is TRANSLOCATED from cytosol to membranes of type II pneumocytes . CNI

Valence Pattern

Seen in 8 of 17 GRIFs with this LU (47.06%)

Transport_destination	Transported_entity	Transporting_entity
PP[to]	NP	CNI
Dep	Ext	--

GRIF 8866

EntrezGene ID: 25400 [symbol: Camk2a]

PubMed Doc ID: 1468861

CaMKII-alpha is TRANSLOCATED to the cell membranes , particularly synaptic membranes , where it may modulate cellular function CNI

GRIF 31033.1

EntrezGene ID(s): NA

PubMed Doc Id(s): NA

NCS-1 protein and 1-phosphatidylinositol 4-kinase b interact in neuronal cells and are TRANSLOCATED to membranes during nucleotide-evoked exocytosis . CNI

GRIF 95249

EntrezGene ID: 24626 [symbol: Pde4b]

PubMed Doc ID: 1582923

From these results , we conclude that either the changes in PDE4B are due to modulation of pre-existing mRNA , or that the protein is specifically TRANSLOCATED to activated synaptic structures . CNI

GRIF 93544

EntrezGene ID: 24182 [symbol: Agtr2]

PubMed Doc ID: 1574609

Constitutively active homo-oligomeric AT2 receptor by intermolecular interaction in two extracellular loops is TRANSLOCATED to the cell membrane and induces cell signaling independent of receptor conformation and ligand stimulation . CNI

GRIF 34624

EntrezGene ID: 24681 [symbol: Prkcc]

PubMed Doc ID: 1468861

PKC is TRANSLOCATED to the cell membranes , particularly synaptic membranes , where it may modulate cellular function CNI

Valence Pattern

Seen in 2 of 17 GRIFs with this LU (11.76%)

Transport_destination	Transported_entity	Transporting_entity
PP[to]	NP	CNI
Dep	Obj	--

GRIF 82174

EntrezGene ID: 66013 [symbol: Arhgef9]

PubMed Doc ID: 1521530

TRANSLOCATES gephyrin to submembrane microaggregates CNI

Valence Pattern

Seen in 3 of 17 GRIFs with this LU (17.65%)

Transport_destination	Transported_entity	Transporting_entity
PP[into]	NP	CNI
Dep	Ext	--

GRIF 47966

EntrezGene ID: 835895 [symbol: AT5G57850]

PubMed Doc ID: 1550046

The full-length Arabidopsis ADC lyase polypeptide was TRANSLOCATED into isolated pea chloroplasts and directed the passenger protein to Arabidopsis chloroplasts . CNI

GRIF 7607

EntrezGene ID: 914 [symbol: CD2]

PubMed Doc ID: 1242637

CD2BP2 is the ligand of the membrane-proximal proline-rich tandem repeat of CD2 in detergent-soluble membrane compartments , but is replaced by Fyn SH3 after CD2 is TRANSLOCATED into lipid rafts upon CD2 ectodomain clustering . CNI

Valence Pattern

Seen in 2 of 17 GRIFs with this LU (11.76%)

Transport_locations	Transported_entity	Transporting_entity
CNI	NP	CNI
--	Obj	--

GRIF 65984.1

EntrezGene ID(s): NA

PubMed Doc Id(s): NA

hexosamine biosynthesis pathway-mediated activation of PI 3-kinase has an insulin-like effect to TRANSLOCATE GLUT4 and causes PI 3-kinase-mediated translocation of PKC-zeta/lambda and PKC-varepsilon but not other PKC isoforms tested (alpha , beta , delta) . CNI CNI

GRIF 76838

EntrezGene ID: 20304 [symbol: Ccl5]

PubMed Doc ID: 1189373

role in chemokines transcription in astrocytes involves activating and TRANSLOCATING p90 ribosomal S6 kinase CNI CNI

Valence Pattern

`Seen in 1 of 17 GRIFs with this LU (5.88%)`

Transport_locations	Transported_entity	Transporting_entity
CNI	NP	CNI
--	Ext	--

GRIF 45884

EntrezGene ID: 5781 [symbol: PTPN11]

PubMed Doc ID: 1255246

The CagA protein of Helicobacter pylori is TRANSLOCATED into epithelial cells and binds to SHP-2 in human gastric mucosa CNI CNI

Nominal lexical unit: Cause_protein_transport.translocation.n

GRIF Annotations (*seen in 3 Valence Patterns*)

Valence Pattern

`Seen in 1 of 3 GRIFs with this LU (33.33%)`

Transport_locations	Transported_entity	Transporting_entity
CNI	PP[of]	Poss
--	Dep	Gen

GRIF 37098

EntrezGene ID: 5580 [symbol: PRKCD]

PubMed Doc ID: 1549452

Protein kinase C delta plays a pivotal role in stimulating monocyte NADPH oxidase activity through its regulation of phosphorylation and TRANSLOCATION of p47phox . CNI

Valence Pattern

`Seen in 1 of 3 GRIFs with this LU (33.33%)`

Transport_locations	Transported_entity	Transporting_entity
CNI	PP[of]	PP[by]
--	Dep	Dep

GRIF 77473

EntrezGene IDs: 24137 [symbol: KIF4A] , 9055 [symbol: PRC1]

PubMed Doc ID: 1562510

role of PRC1 in midzone formation , indicate that cell cycle-dependent TRANSLOCATION of PRC1 by Kif4 is essential for midzone formation and cytokinesis . CNI

Valence Pattern

Seen in 1 of 3 GRIFs with this LU (33.33%)

Transport_locations	Transported_entity	Transporting_entity
CNI	CNI	PP[by]
--	--	Dep

GRIF 76067

EntrezGene ID: 950614 [symbol: hslU]

PubMed Doc ID: 1503724

results suggest a similarity in mechanism of the apparent rate-limiting steps of unfolding and TRANSLOCATION by the chaperone components HslU and ClpX CNI CNI

Appendix 3: Predicates in HLKB Transport Classes and BioFN Frames

Lexical Units' Presence in BioOntology Classes, presented in context of BioFrames

(Protein_transport, Cause_protein_transport)

	Protein transport [Prot_transp]	Protein transport [Cause_prot_transp]	Gated_nucl transport [Prot_transp]	Gated_nucl transport [Cause_prot_transp]	Transmembr transport [Prot_transp]	Transmembr transport [Cause_prot_transp]	Vesic transport [Prot_transp]	Vesic transport [Cause_prot_transp]	Endocyt [Prot_transp]	Endocyt [Cause_prot_transp]
anchor.v						1				
delivery.n			1							
distribute.v				1						
efflux.n			1							
endocytosis.n									76	
enter.v			1							
entry.n			1							
exclude.v				1						
exit.v			1							
exocytosis.n							2			
export.n		1	52	1						
export.v				6						
import.n			37							
import.v				1		1				
internalization.n									8	
internalize.v										2
migrate.v			1							
mobilization.n	1		1							
move.v			3		2					
movement.n			2							
recruit.v				1						
recruitment.n					1					
recycle.v		1					3	1		

	Protein transport [Prot_transp]	Protein transport [Cause_prot_transp]	Gated_nucl transport [Prot_transp]	Gated_nucl transport [Cause_prot_transp]	Transmembr transport [Prot_transp]	Transmembr transport [Cause_prot_transp]	Vesic transport [Prot_transp]	Vesic transport [Cause_prot_transp]	Endocyt [Prot_transp]	Endocyt [Cause_prot_transp]
recycling.n	1									
redistribution.n	2		2		1		1			
release.n	4				14					
release.v						3				
relocalize.v				1						
relocate.v			1					1		
relocation.n	1									
return.v			1							
sequester.v				1		1				
shift.n			1							
shuttle.v			15							
shuttling.n			4							
sort.v								1		
target.v		2		3		10				1
targeting.n			2		2		1			
traffic.n							4			
trafficking.n			1				13		3	
translocate.v	10	17	17	9	4	2	2	1	1	
translocation.n	207	3	185	4	32		4		1	
transport.n	4		19		1		10			
transport.v			1	2				1		

Appendix 4: Lexical Entry Reports for 'translocation.n'

translocation.n

Frame: Protein_transport

Definition

This frame involves the phenomenon of intracellular transport, the directed movement within a cell of a Transported_entity from a Transport_origin to a different location, the Transport_destination. Alternatively, Transport_locations may be mentioned with no specific indication of origin vs. destination, or the location is both origin and destination in continuous, frequent motion events. Movement of the Transported_entity follows one of a variety of transport mechanisms; these may involve crossing specific membranes, moving through pores within a subcellular component such as the nucleus, or making use of membrane-enclosed transport intermediates.

Frame Elements and Their Syntactic Realizations

The Frame elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realization(s)
Transport_destination	(75)	A.Dep (4) PP[to].Dep (56) PP[onto].Dep (1) PP[into].Dep (6) N.Dep (8)
Transport_locations	(132)	PP[across].Dep (2) PP[in].Dep (1) CNI.-- (126) PP[through].Dep (1) PP[at].Dep (2)
Transport_origin	(7)	PP[from].Dep (5) CNI.-- (1) --.-- (1)
Transported_entity	(207)	NP.Ext (3) N.Dep (105) PP[of].Dep (79) Poss.Gen (10) CNI.-- (10)

Valence Patterns:

These frame elements occur in the following syntactic patterns:

Number Annotated	Patterns		
7 TOTAL	Transport_destination	Transport_origin	Transported_entity
(3)	PP[to] Dep	PP[from] Dep	PP[of] Dep
(2)	PP[to] Dep	PP[from] Dep	N Dep
(1)	PP[to] Dep	CNI --	Poss Gen
(1)	PP[to] Dep	-- --	PP[of] Dep
68 TOTAL	Transport_destination	Transported_entity	
(7)	N Dep	PP[of] Dep	
(2)	PP[to] Dep	NP Ext	
(15)	PP[to] Dep	N Dep	
(3)	PP[into] Dep	PP[of] Dep	
(1)	PP[into] Dep	Poss Gen	
(3)	PP[to] Dep	Poss Gen	
(2)	PP[into] Dep	N Dep	
(24)	PP[to] Dep	PP[of] Dep	
(1)	PP[onto] Dep	PP[of] Dep	
(1)	N Dep	N Dep	
(2)	A Dep	PP[of] Dep	
(1)	A Dep	NP Ext	
(1)	A Dep	CNI --	
(5)	PP[to] Dep	CNI --	

132 TOTAL	Transport_locations	Transported_entity	
(1)	PP[through] Dep	N Dep	
(2)	PP[across] Dep	PP[of] Dep	
(1)	PP[in] Dep	N Dep	
(2)	PP[at] Dep	PP[of] Dep	
(34)	CNI --	PP[of] Dep	
(5)	CNI --	Poss Gen	
(83)	CNI --	N Dep	
(4)	CNI --	CNI --	

translocation.n

Frame: Cause_protein_transport

Definition

A Transporting_entity, typically a protein, mediates the transport of a Transported_entity, a different molecular entity, within a cell from a Transport_origin to a Transport_destination, a different location. Alternatively, Transport_locations may be mentioned with no specific indication of origin vs. destination, or the location is both origin and destination in continuous, frequent motion events. Movement of the Transported_entity follows one of a variety of transport mechanisms; these may involve crossing specific membranes, moving through pores within a subcellular component such as the nucleus, or making use of membrane-enclosed transport intermediates.

Frame Elements and Their Syntactic Realizations

The Frame elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realization(s)
Transport_locations	(3)	CNI.-- (3)
Transported_entity	(3)	PP[of].Dep (2) CNI.-- (1)
Transporting_entity	(3)	PP[by].Dep (2) Poss.Gen (1)

Valence Patterns:

These frame elements occur in the following syntactic patterns:

Number Annotated	Patterns		
3 TOTAL	Transport_locations	Transported_entity	Transporting_entity
(1)	CNI --	PP[of] Dep	Poss Gen
(1)	CNI --	PP[of] Dep	PP[by] Dep
(1)	CNI --	CNI --	PP[by] Dep