Taylor & Francis
Taylor & Francis Group

Natural language generation of biomedical argumentation for lay audiences

Nancy Green[a]*, Rachael Dwight[b], Kanyamas Navoraphan[b] and Brian Stadler[a]

[a]*Computer Science Department, University of North Carolina Greensboro, Greensboro, NC 27402, USA;*
[b]*Computer Science Department, North Carolina State University, Raleigh, NC, USA*

This article presents an architecture for natural language generation of biomedical argumentation. The goal is to reconstruct the normative arguments that a domain expert would provide, in a manner that is transparent to a lay audience. *Transparency* means that an argument's structure and functional components are accessible to its audience. Transparency is necessary before an audience can fully comprehend, evaluate or challenge an argument, or re-evaluate it in light of new findings about the case or changes in scientific knowledge. The architecture has been implemented and evaluated in the *Gen*etics *I*nformation *E*xpression Assistant, a prototype system for drafting genetic counselling patient letters. Argument generation makes use of abstract argumentation schemes. Derived from the analysis of arguments used in genetic counselling, these mainly causal argument patterns refer to abstract properties of qualitative causal domain models.

Keywords: natural language generation; computational models of natural argument; biomedical argumentation; causal argumentation schemes

1. Introduction

Healthcare consumers face increasing responsibility for assimilating the information given to them by healthcare providers (Nielsen-Bohlman, Panzer, Hamlin, and Kindig 2004). There are many reasons why understanding this information can be challenging. Perhaps the most obvious is that the information may contain specialised biomedical terminology or presuppose certain background knowledge. A related problem is the public's low level of numeracy (Ancker and Kaufman 2007). Beyond having biomedical knowledge and numeracy skills, the audience must be able to fully comprehend the arguments for the diagnosis and other tentative conclusions of the healthcare provider. That requires, for example, the ability to identify an argument's premises, which may be explicit or implicit in a text. An argument is *transparent* when its structure and functional components are accessible to its audience. Transparency is necessary before an audience can fully comprehend, evaluate or challenge an argument, or re-evaluate it in light of new findings about the patient or changes in scientific knowledge. Because of the importance of argument comprehension for healthcare consumers, our research is on the natural language generation (NLG) of transparent biomedical argumentation.

As a testbed for the research, we developed the GenIE (*Gen*etics *I*nformation *E*xpression) Assistant, a prototype system that generates the first draft of a genetic counselling *patient letter*. This letter is a standard document written by a genetic counsellor to her client summarising information and services provided to a patient (Baker, Eash, Schuette, and Uhlmann 2002). It is written after the counsellor meets with a client and is not intended to serve as the principal means of communication with the client. A preliminary analysis of a corpus of patient letters showed that they contain a variety of argument patterns[1] (Green 2007). Thus, this genre can provide insight

---

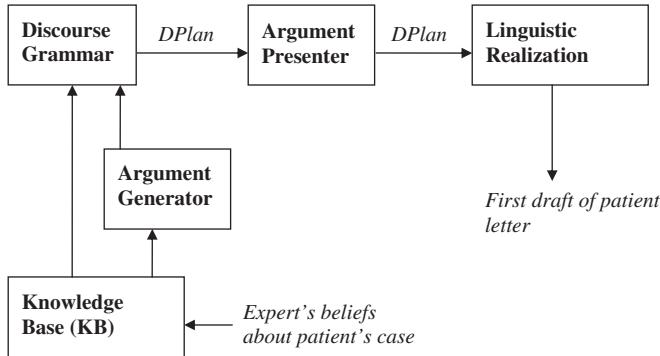*Corresponding author. Email: nlgreen@uncg.edu

Figure 1. Architecture of GenIE Assistant.

into biomedical argumentation designed by domain communication experts for a lay audience. A practical use of a system such as the GenIE Assistant would be as an authoring tool for genetic counsellors.[2]

The components of the GenIE Assistant are shown in Figure 1. A conceptual model of clinical genetics and information on a specific patient are represented in a *knowledge base* (KB). The generation process begins using the *discourse grammar* to construct a partial *discourse plan* (DPlan). Discourse grammar rules encode genre-specific knowledge of the standard topical organisation of patient letters and extract basic information about a patient's case from the KB. As in many other NLG systems, a DPlan represents the structure and content of the text to be generated in non-linguistic form (Reiter and Dale 2000). In the GenIE Assistant, the structure is described in terms of the rhetorical structure theory (RST) relations (Mann and Thompson 1988). However, the discourse grammar is not used to generate arguments, which are produced by GenIE's *argument generator*.

As a DPlan is constructed by the discourse grammar, claims requiring support are sent by the discourse grammar to the argument generator. The argument generator produces arguments for a given claim by instantiating argumentation schemes, abstractions of normative[3] arguments found in the corpus, with information from the KB. For the sake of generality, the argumentation schemes refer to abstract properties of a KB rather than to genetics. Complex arguments containing subarguments may be constructed by composing instantiated argumentation schemes. The resulting argument structures are translated by the argument generator into equivalent RST structures, which are inserted into the DPlan. The responsibility of the *argument presenter* is to make changes in the DPlan for the sake of coherence and transparency of arguments in the generated text. In the final phase of processing, *linguistic realisation*, NLG microplanning, and surface realisation tasks (Reiter and Dale 2000) are done to render the DPlan as English text.

One of the main contributions of this article is the above software architecture, the design of which was informed by a corpus study of how domain communication experts present biomedical arguments in writing to a lay audience. The architecture has been implemented in a prototype system – the GenIE Assistant, and the KBs and argument generator have been reused in another prototype system – Interactive GenIE. Due to the separation of domain content from argumentation concerns, the design should be portable to other domains and applications.

The other main contribution of this work is the definition of causal argumentation schemes and their critical questions in terms of abstract properties of qualitative probabilistic/causal domain models. A critical question is a question associated with an argumentation scheme such that its answer may defeat the defeasible claim of an argument constructed from the scheme (Walton,

1. This letter will summarize the testing performed on blood obtained from [P] to determine the cause of [P's] hearing loss.
2. We met with [CLIENT] to discuss these results in [CLINIC] on [DATE].
3. [P's] blood was obtained to determine if [P's] hearing loss was due to a change in a gene called GJB2 (for gap junction protein beta 2) that is associated with childhood hearing loss.

…

8. The most common change in this gene causing deafness is called 35delG and involves the loss (deletion) of a small piece of genetic material within the gene.
9. The change found in [P's] GJB2 gene was a rarer change and is called [MUTATION].
10. This change is known to result in a shortened or truncated protein.
11. The protein, called Connexin 26, is important in maintaining the chemical equilibrium of the inner ear.

Figure 2. Excerpt from patient letter in corpus.

Reed, and Macagno 2008). Computational use of argumentation schemes and critical questions is an active research area, e.g. in artificial intelligence and law (Gordon, Prakken, and Walton 2007), artificial intelligence and medicine (Lindgren and Eklund 2005; Fox et al. 2007; Rahati and Kabanza 2009), and bioinformatics (McLeod and Burger 2008). However, that work has not addressed the argumentation based upon domain knowledge represented in causal models. Furthermore, this article helps in bridging the gap between research on computational use of argumentation schemes and research on NLG.

2. Corpus analysis

This section motivates the GenIE Assistant's design by giving an overview of how it has been informed by qualitative analysis of different aspects of a corpus of 21 patient letters written by genetic counsellors. Consistent with professional guidelines (Baker et al. 2002), the letters typically include a patient's history (including symptoms, a preliminary diagnosis, and any testing to confirm the preliminary diagnosis and the results of testing), a final diagnosis, the probable source of inheritance of the genetic condition, and the patient's family members' inheritance risks. The model presented in this paper is based mainly on nine letters in the corpus that cover nine different patients and seven different single-gene autosomal genetic disorders,[4] and come from four different authors (each at a different institution). The letters range in length from 24 sentences (446 words) to 76 sentences (1537 words). For illustration, see an excerpt from one letter in Figure 2. In the figure, we have numbered the sentences for ease of reference, and to maintain client privacy, some words have been replaced by placeholders enclosed in square brackets; e.g. (*P*) refers to the patient.

2.1. *Analysis of conceptual model of domain*

On initial review of the corpus, it was apparent that certain recurrent general concepts are used to convey a simple, mainly causal model of genetic inheritance and disease to its lay audience. For example, the text of Figure 2 can be analysed as referring to the following concepts (italicised): *test* ('testing'), *symptom* ('hearing loss'), *test result* ('these results'), *genotype* ('GJB2'), and *biochemistry* ('Connexin 26'). A formal evaluation, reported in (Green 2005a), was performed to validate the inter-coder reliability of a set of eight concepts (listed in Table 1) for letters on single-factor autosomal genetic conditions. Using instances of these concepts, it is possible to manually reconstruct qualitative causal/probabilistic network graphs representing the domain content of a letter.[5] The nodes of the graph represent the instances of concepts referred to in a letter (e.g. the patient's GJB2 genotype), while node types (e.g. *genotype*) correspond to the more abstract concepts shown in Table 1.

Table 1. Concepts in genetic counselling corpus.

Concept	Description	Example
History	Demographic or other risk factor	Ethnicity
Genotype	Number (0, 1, 2) of mutated alleles of a gene	Two mutated alleles of GJB2
Event	External or mutagenic event	Presence of bacteria
Biochemistry	Manifestation of genotype at biochemical level	Abnormal Connexin 26 protein
Physiology	Manifestation of genotype at physiological level	Malabsorption
Symptom	Observable symptom	Hearing loss
Test	Decision to perform a test	Sweat test performed
Test result	Result of test	Abnormal NaCl level

2.2. *Analysis of argumentation*

In addition to analysing the conceptual domain model underlying the letters, we analysed the argumentation schemes used in the letters. It is important to note that a challenge to identification of the argumentation schemes was that the arguments often require fine-grained discourse interpretation, including reconstruction of implicit premises and/or conclusions (Green 2010). For generality, the argumentation schemes were defined in terms of abstract properties of the conceptual domain models rather than domain-specific concepts. To support transparency, the premises of an argument were identified by function as data or warrant (Toulmin 1998), where the warrant is typically a defeasible generalisation linking data to claim.

For illustration, consider the argument partly expressed in sentences 10 and 11 of Figure 2. The implicit *claim* of this argument is that, in each of *P*'s cells, both copies (alleles) of the GJB2 gene have the mutation referred to as (MUTATION). The implicit *data* are that the patient has hearing loss. Note that the data were not provided explicitly since, not only would it be shared knowledge of the genetic counsellor and her client, but it is mentioned in the preceding text. The *warrant* is the causal path from the GJB2 genotype to hearing loss represented in the conceptual model of the domain. Sentences 10 and 11 describe the causal chain up to chemical equilibrium of the inner ear, but the path from that to hearing loss is only implicated. As this example illustrates, a writer may rely on a reader's ability to reconstruct implicit components of an argument. For someone who is not a domain expert, or skilled in processing argumentation, reconstruction of arguments could be a difficult task.

2.3. *Analysis of discourse*

Several aspects of discourse in the corpus were analysed. First, an RST analysis of several letters was performed. Along with published guidelines for writers of patient letters (Baker et al. 2002), the RST analysis informed the specification of the discourse grammar.[6] Second, an analysis of three dimensions of arguments in four letters was done to inform argument presentation and linguistic realisation: order of components (claim, data, warrant), implicit components, and use of discourse cues associated with argument components. The results of that study are summarised in Section 5.1. The next three sections present the major components of the GenIE architecture.

3.   Knowledge base

3.1. *Conceptual domain model*

The KB represents the counsellor's beliefs about the domain and, before a letter is generated on a particular patient's case, must be updated with his beliefs about the case (findings, diagnosis, etc.). As described in Section 2.1, a corpus study was carried out to identify the set of eight recurrent

general concepts in this genre (Table 1). The domain content of letters to be generated can be represented as network graphs whose nodes are instances of these eight general concepts. Arcs in the graph represent qualitative causal/probabilistic relations determined by the types of the nodes connected by the arcs. The design of domain models for GenIE is restricted to this set of node types and relations.
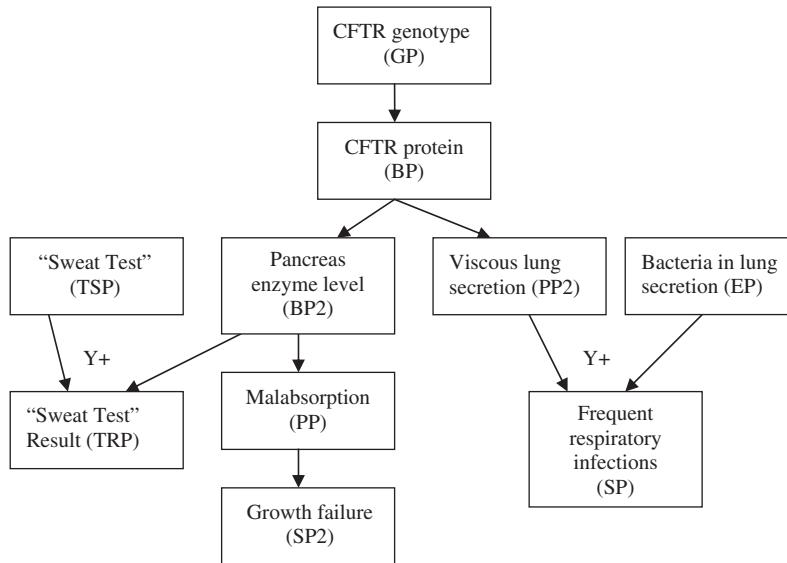
Restricting the models in this way has several benefits. First, it simplifies the computational process of selecting the content for arguments, since the generator does not have to filter out biomedical information that is not normally used in communication with a lay audience. Second, it guides domain knowledge acquisition since it is constrained by the set of available concepts and relations. Using this set of node types and relations, prototype KBs for a number of representative genetic conditions[7] were manually implemented,[8] each in little time. Third, this approach enables argumentation schemes derived from the analysis of the corpus to be specified in terms of abstract properties of a domain model, as will be shown in Section 4.

3.2. *Computational representation*

A KB is modelled computationally as a qualitative probabilistic network (QPN) (Wellman 1990; Druzdzel and Henrion 1993). Like a Bayesian network (Korb and Nicholson 2004), a QPN is a directed acyclic graph whose nodes represent random variables. However, in a QPN, qualitative constraints take the place of conditional probability tables. The qualitative constraints are defined in terms of the relations of qualitative influence S, additive synergy Y, and product synergy X.[9] A variable A is said to have a positive qualitative influence on a variable B, written $S^+(A, B)$, if a higher value of A makes a higher value of B more likely. An analogous definition is given for negative influence, i.e. $S^-(A, B)$, if a higher value of A makes a lower value of B more likely. Each arc in a KB is described by a positive or negative influence relation. Since not all the variables are Boolean, our notation encodes threshold values of the domains explicitly. For example, positive influence is encoded as $S^+(<A, a>, <B, b>)$, i.e. when A reaches a (or higher), it is more likely that B will reach b (or higher).

Synergy relations X and Y are used to express a relationship between a set of variables and another variable. Negative product synergy, $X^-(\{A, B\}, <C, c>)$, is used to model the relation between a mutually exclusive set $\{A, B\}$ of potential causes of an event $<C, c>$ and the event, e.g. between a set of candidate diagnoses and presence of a certain symptom. Another example of its use in this domain is to model autosomal dominant inheritance; having one mutated allele of a gene fitting this inheritance pattern normally is sufficient to cause a health problem. Thus, if a child has an autosomal dominantly inherited disorder, then (since we inherit one copy of each pair of genes from each parent) the child must have inherited the mutation from one or the other of his parents. Zero product synergy, $X^0(\{A, B\}, <C, c>)$, is used to model the relation between a set of events $\{A, B\}$ which are jointly necessary to cause the event $<C, c>$; e.g. in autosomal recessive inheritance of a disorder, the child must have inherited one copy of the mutation from each parent. Positive additive synergy, $Y^+(\{A, B\}, <C, c>)$, is used to model a situation where A enables B to result in C. Negative additive synergy, $Y^-(\{A, B\}, \neg<C, c>\}$, where the notation $\neg<C, c>$ denotes $C < c$, is used to model a situation in which A inhibits B from resulting in C.

Figure 3 shows part of a KB describing a patient believed to have cystic fibrosis (CF) based on his symptoms – growth failure and respiratory infections – as well as the test result showing an abnormal level of NaCl. All arcs in the figure are represented in the KB as positive influence relations (S^+). In addition, the graph contains converging arcs, annotated Y^+, which are represented as two additive synergy relations in the KB: $Y^+(\{TSP, BP2\}, <TRP, abnormal>)$ and $Y^+(\{EP, PP2\}, <SP, true>)$. The first one can be glossed as: *performing the test* (TSP)

Figure 3. Part of KB for CF case. Concept types are from the generic model of genetic disease and inheritance. Variables describe (part of) the model of CF. States of the variables describe (part of) a particular patient's case.

enables the pancreas enzyme level (BP2) *to be detected in the test result* (TRP). The second one can be glossed as: *the presence of bacteria* (EP) *enables the viscous lung secretions* (PP2) *to lead to frequent respiratory infections* (SP). Note that the node types and possible relations among nodes are determined by the general conceptual model of genetic disease; the graph instantiates this model for CF; and the values of the variables, given below the graph, describe some of the expert's beliefs about a particular patient's case after the patient's test result is known.

Figure 4 shows another part of the same KB, focusing on the conceptual model of the probable source of the patient's inherited genetic disease. This part describes the expert's beliefs that the patient inherited two mutated alleles of the CFTR genotype, one from each parent, but that each parent carries only one copy of the mutation since neither shows symptoms of CF. Note that, for each parent, a history variable representing North European ancestry is used for representing an individual's higher risk of carrying one mutated CFTR allele in that population compared with other groups.

Figure 4. Another part of KB for CF case. This part shows information relevant to claims about the patient's parents' CFTR genotype.

4. Argument generation

In the GenIE Assistant, argumentation for a given claim is reconstructed by instantiating argumentation schemes with information from the KB. The argumentation schemes currently implemented in the GenIE Assistant are defined in Section 4.1. An example of an argument composed from several of the argumentation schemes is given in Section 4.2, followed by a description of the generation process in Section 4.3. In Section 4.4, a model of interactive argument generation is presented.

4.1. *Argumentation schemes*

As a prerequisite to identification of argumentation schemes, we interpreted the conceptual model conveyed in a letter as it would be represented in a KB (as described in Section 3). Next, we analysed arguments in the letter in terms of their claim, data, and warrant.[10] In some cases, it was necessary to interpret the writer's intended message, constrained by the conceptual model of the domain, the discourse context, and knowledge presumed to be shared by writer and audience to reconstruct implicit components of an argument (Green 2010). The final step in deriving an argumentation scheme was to abstract from the terms of a specific conceptual model and represent the claim, data, and warrant in terms of variables and abstract properties (i.e. influence and synergy relations) of a KB.

   To illustrate, an argumentation scheme describing the argument analysed in Section 2 is shown at the top of Table 2. The warrant of the argument is that there is a positive influence relation, or a chain of such influences, from A to B. (To simplify the presentation, a chain is denoted as $S^*(<A, a>, <B, b>)$, i.e. S^* is used to denote positive influence along a path of length one

Table 2. Effect to cause argumentation schemes.

Components of scheme	Example
Simpler case	
Claim: $A \geq a$	P has two mutated GJB2 alleles
Data: $B \geq b$	P has hearing loss
Warrant: $S^*(<A, a>, <B, b>)$	Having a GJB2 genotype with two mutated copies can result in ... which can result in hearing loss
Applicability constraint: $\neg \exists C X^-(\{C, A\}, <B, b>)$: $C \geq c$	Unless something else is thought to be the cause of the hearing loss
Variant	
Claim: $A \geq a$	P has two mutated CFTR alleles
Data: $B \geq b$ & $D \geq d$	P has frequent respiratory infections and bacteria in lung secretions
Warrant: $S^*(<A, a>, <B, b>)$ & $Y^+(\{D, A'\}, <B, b>)$, where A' is on path from A to B	Having a CFTR genotype with two mutated copies can result in ... viscous lung secretions which, enabled by the presence of bacteria, can result in respiratory infections
Applicability constraint: $\neg \exists C X^-(\{C, A\}, <B, b>)$: $C \geq c$	Unless something else is thought to be the cause of the respiratory infections

or more.) The data are that B has reached b. The claim (conclusion) is that A has reached a.[11] Of course, this is not a deductively valid conclusion. For instance, there could be an alternative explanation for B reaching b, if there were another variable C such that C reaching the value c would increase the probability that B reaches the value b.

Since, ideally, a writer would not use an argumentation scheme that he did not believe to be applicable, and to prevent the GenIE Assistant from constructing arguments that would not be constructed by the writer, an *applicability constraint* may be included in an argumentation scheme. For an argumentation scheme to be applicable, its applicability constraint must hold. To gloss the applicability constraint of the scheme described above, it says that there is no variable C in a negative product synergy relation $X^-(\{C, A\}, <B, b>)$ such that C has reached c. A violation of this applicability constraint is depicted in Figure 5(a). Note that the description of the applicability
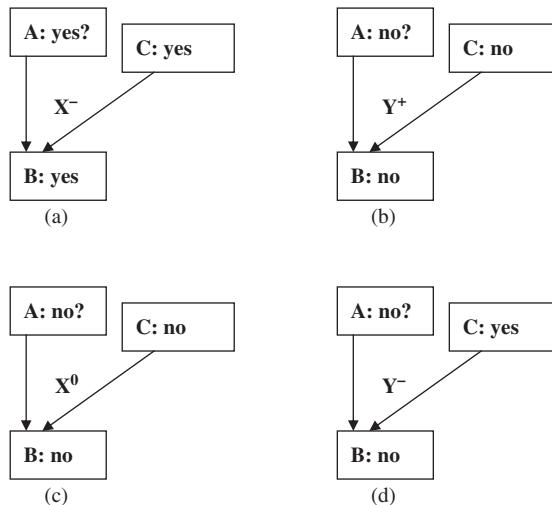

Figure 5. (a) C could be the explanation for B. (b) C could have failed to enable A to influence B. (c) C could have failed to jointly contribute to B. (d) C could have prevented B.

constraint actually implemented for warrants with a path length greater than one is not shown; it is more complicated, since one must consider alternative explanations for intermediate steps in the path from A to B.

In general, the applicability constraint of an argumentation scheme was derived by consideration of qualitative constraints that could be used to encode a KB (as described in Section 3) and that represented potential exceptions to the applicability of the argumentation scheme. As will be discussed in Section 4.4, each condition of the applicability constraint can be thought of as a critical question that can be used to challenge the conclusion.[12]

We now discuss the remaining argumentation schemes implemented in this version of the GenIE Assistant, followed by an example of how a typical argument can be composed from instances of several of the schemes.

A variant of the argumentation scheme described above is shown at the bottom of Table 2. For example, the following argument can be constructed from it: given the beliefs (as shown in Figure 3) that (*data*) the patient has had frequent respiratory infections and bacteria in his lung secretions, and that (*warrant*) having two mutated alleles of CFTR leads to viscous lung secretions and the presence of bacteria enables viscous lung secretions to lead to frequent respiratory infections, then (*conclusion*) the patient has two mutated alleles of CFTR.

Another argumentation scheme is shown in Table 3. For example, as warranted by an autosomal recessive inheritance pattern, the data that a child has two mutated alleles of a gene support the claim that each parent has at least one mutated allele of that gene.

In contrast to the above argumentation schemes, consider the argumentation scheme shown in Table 4. The warrant of the argument is identical to that given in Table 2. However, the data are that B has *not* reached b, and the claim (conclusion) is that A has *not* reached a. For example, lack of symptoms is evidence for the claim that an individual does *not* have the genetic condition associated with those symptoms. The applicability constraint of this scheme has three conditions, all of which must hold.

The first condition states that there is no variable C and positive additive synergy relation, $Y^+(\{C, A\}, <B, b>)$, such that C did *not* reach the value c; i.e. the explanation for B not reaching b could be that by not reaching c, C failed to enable A to influence B to reach b (Figure 5(b)). For example, if exposure to bacteria (C) is required to enable thickened mucous in the lungs (A) to

Table 3. Effect to joint cause argumentation scheme.

Components of scheme	Example
Claim: $(A \geq a)$ and $(C \geq c)$	P's mother and father each have at least one mutated allele of GJB2
Data: $B \geq b$	P has two mutated alleles of GJB2
Warrant: $X^0(\{A, C\}, <B, b>)$	A child inherits one allele of GJB2 from each parent

Table 4. No effect to no cause argumentation scheme.

Components of scheme	Example
Claim: $A < a$	P does not have CF
Data: $B < b$	P does not have symptoms of CF
Warrant: $S^*(<A, a>, <B, b>)$	Someone with CF exhibits symptoms of CF
Applicability constraint: $[\neg\exists C : Y^+(\{C, A\}, <B, b>) : C < c]$ and $[\neg\exists C : X^0(\{C, A\}, <B, b>) : C < c]$ and $[\neg\exists C : Y^-(\{C, A\}, \neg<B, b>) : C \geq c]$	

lead to respiratory infections (B), then an argument for the absence of thickened mucous in the lungs based on the absence of respiratory infections is not applicable unless it is plausible that there *was* exposure to bacteria.

The second condition states that there is no variable C and zero product synergy relation, $X^0(\{C, A\}, <B, b>)$, such that C did not reach c; i.e. the explanation could be that A and C are jointly required and the reason B failed to reach b was that C did not reach c (Figure 5(c)). For example, in an autosomal recessive inheritance pattern, a child with two inherited mutations of a gene received one mutated allele from each parent; thus, if the child does not have two inherited mutated alleles of a gene and one parent is believed to lack the mutation, then an argument for the absence of the mutation in the other parent based on the child's not having two inherited mutations is not applicable.

The third condition states that there is no variable C and negative additive synergy relation, $Y^-(\{C, A\}, \neg<B, b>)$, such that C did reach c; i.e. the explanation could be that by reaching c, C prevented A from influencing B to reach b (Figure 5(d)). For example, if a patient is taking medication to treat a possible genetic condition, then failure to manifest the condition's symptoms does not support the claim that the patient does not have the condition.

The argumentation schemes shown in Tables 5 and 6 do not have warrants from the domain model as they represent commonsense reasoning patterns.

Another argumentation scheme identified in the corpus is shown in Table 7. For example, an autosomal dominant inheritance pattern fits the warrant of this scheme. Thus, the data that the child has exactly one mutated allele of a gene support the claim that either the mother or father (but not both) has at least one mutated allele of that gene. In a variant of this scheme (not shown in Table 7) found in the corpus and implemented in the GenIE Assistant, the warrant describes three mutually exclusive events.

While most of the argumentation schemes discussed so far describe arguments from effects (or lack thereof) to potential causes, the next argumentation scheme, shown in Table 8, uses as

Table 5. Conjunction simplification argumentation scheme.

Components of scheme	Example
Claim: P	The mother has one or two mutated alleles of GJB2
Data: P *and* Q	The mother has one or two mutated alleles of GJB2, and so does the father

Table 6. Elimination argumentation scheme.

Components of scheme	Example
Claim: $P = p_1$	The mother has exactly one mutated allele of GJB2
Data: $[(P = p_1)$ or $(P = p_2)]$ and $(P \neq p_2)$	The mother has one or two mutated alleles of GJB2, and she does not have two of them

Table 7. Effect to Alternative Causes argumentation scheme.

Components of scheme	Example
Claim: $(A = a)$ or $(C = c)$	Either P's mother or P's father has exactly one mutated allele of FGFR3
Data: $B = b$	P has one mutated allele of FGFR3
Warrant: $X^-(\{A, C\}, <B, b>)$	A child inherits one mutated allele of FGFR3 from one or the other of his parents

Table 8. Increased risk argumentation scheme.

Components of scheme	Example
Claim: $B \geq b$	P will have heart disease
Data: $A \geq a$	P has FH
Warrant: $S^*(<A, a>, <B, b>)$	Having FH increases the risk of heart disease
Applicability constraint:	Unless:
$[\neg \exists C : Y^-(\{C, A\}, \neg <B, b>) :$	(1) one is taking preventive medication, and
$C \geq c]$ and $[\neg \exists C :$	(2) one has no other risk factors (e.g. obesity)
$Y^+(\{C, A\}, <B, b>) : C < c]$	

data an event that may either play a causal role in or signal risk of another event. For example, obesity is a risk factor for heart disease.

The first condition of the applicability constraint of the scheme in Table 8 is that there is no variable C and negative additive synergy relation, $Y^-(\{C, A\}, \neg <B, b>)$, such that C did reach c; i.e. C may prevent A from influencing B to reach b. For example, medication may lower a risk factor for heart disease. The second condition is that there is no variable C and positive additive synergy relation, $Y^+(\{C, A\}, <B, b>)$, such that C did not reach c; i.e. C may fail to enable A to influence B to reach the value b. For example, the risk of heart disease caused by an inherited condition, familial hypercholesterolemia (FH), can be mitigated by reducing risk factors such as obesity. A related argumentation scheme (not shown) can be used to argue *against* the claim that B has reached (or will reach) b, given the same warrant as in Table 8, and the data that A has *not* reached a.

In summary, the argumentation schemes implemented in this version of the GenIE Assistant are those shown in Tables 2–8 and their variants discussed above. The schemes fall into three categories: an argument for/against a putative cause based upon the observation (or inference) of the presence/absence of its typical effect (Tables 2–4 and 7), an argument for/against the occurrence of a typical effect based upon observation (or inference) of the presence/absence of a causal or risk factor (Table 8), and common sense arguments (Tables 5 and 6).

4.2. *Example of composition of argumentation schemes*

Using the argumentation schemes shown in Tables 2–6, it is possible for the GenIE Assistant to compose a type of argument that is common in genetic counselling, namely, that both biological parents of a child who has an autosomal recessive disorder such as CF are carriers of the mutation responsible for the disorder. (The term *carrier* describes an individual with one normal and one mutated allele of a gene.) A KB representing this situation is shown in Figure 4. The argument for each parent is similar. The argument about the mother of a child with CF is diagrammed in Figure 6 in a notation that highlights the argument's structure and which argumentation schemes were used. For clarity, only data and claims are shown in the diagram; i.e. warrants are not shown.

The main claim that the mother is a carrier (has one mutated CFTR allele) is shown at the top of the diagram. The argument for that claim uses the Elimination argumentation scheme (Table 6). The argument has two premises: (1) that the mother does not have two mutated CFTR alleles and (2) that the mother has at least one mutated CFTR allele. The subargument for premise (1) uses the no effect to no cause scheme (Table 4), and is based on the evidence that the mother does not have symptoms of CF. The subargument for premise (2) uses the conjunction simplification scheme (Table 5), and is based on the claim that the both the mother and the father have at least one mutated CFTR allele. The subargument for that claim uses the effect to joint cause scheme (Table 3), and is based on the belief that the child has CF, i.e. two mutated CFTR alleles. Although not shown in the figure, arguments for the claim that the child has CF can be generated as well.
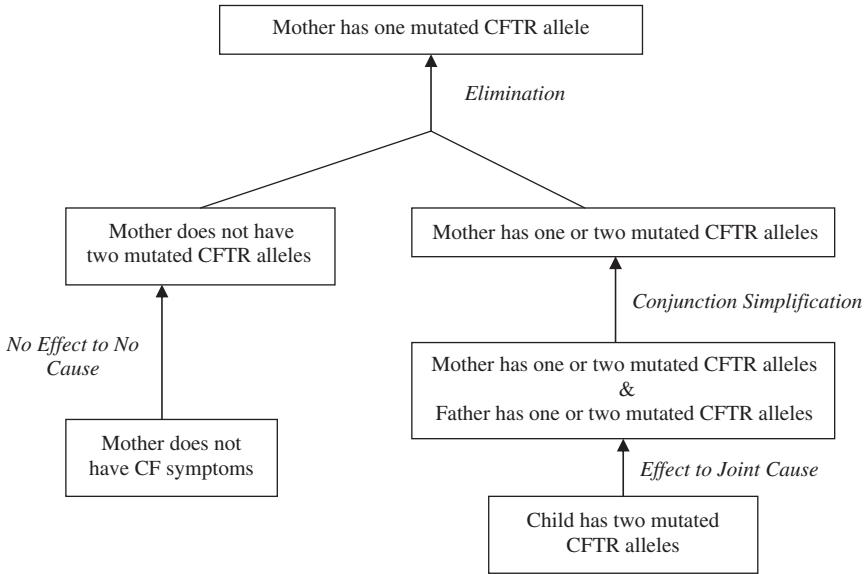
Figure 6. Composition of arguments.

4.3. *Argument generation*

Given a request for all arguments for a given claim, the argument generator processes the request in two phases. The first phase, described in Figure 7, reconstructs all possible arguments for the claim. For example, given the goal to find an argument for the claim that the patient has CF (two mutated copies of the CFTR genotype), the argument generator would find that the claim of the effect to cause argumentation scheme (top of Table 2) can be unified with this goal; next that the warrant of the scheme can be unified with a chain of causal influences in the KB from this genotype to growth failure; next that the data of the scheme can be unified with the assertion in the KB that the patient has growth failure; and lastly that the applicability condition of this scheme holds, i.e. that there is no other condition believed to be applicable in this case. Next, another argument using the variant of the effect to cause scheme (bottom of Table 2) would be found.

In the second phase, the arguments created in the first phase are translated into equivalent RST structures, which are inserted into the discourse structure for the whole document, as follows. The data and claim of an argument are transformed into the satellite and nucleus, respectively, of an evidence relation. The warrant of an argument is transformed into the satellite of a background relation whose nucleus is an evidence relation representing the data and claim of the argument. The RST structure for two arguments for the claim that the patient has CF is shown in Figure 8.

4.4. *Interactive argumentation*

Interactive GenIE (Green 2008), a prototype interactive argument generation system, was implemented to enable a user to take the initiative in exploration of the space of arguments and counterarguments for a given claim. Interactive GenIE consists of a query-driven user interface, the same type of KB used by the GenIE Assistant, and an argument generator. Note that requests for arguments for a claim come from the user interface rather than from the discourse grammar as in the GenIE Assistant. As it was beyond the scope of this project to develop a natural language dialogue interface, communication between user and system is via an artificial language.

```
Input: claim C
Output: set S of all arguments for C

if S exists (i.e. all arguments for C have already been found) return S
if arguments for C have not yet been requested
{
        for each argumentation scheme Arg_i with components: claim C_i, warrant W_i,
        applicability constraint Ap_i, and data D_i
        {
                try unifying C_i  with C, resulting in C_i′
                if successful
                {
                        try instantiating W_i consistent with C_i′ and  KB, resulting in  W_i′
                        if successful
                        {
                                try instantiating D_i consistent with W_i′, resulting in D_i′
                                if successful and if D_i′ holds in KB
                                {
                                        try instantiating Ap_i consistent with W_i′, D_i′
                                        and C_i′, resulting in Ap_i′
                                        if successful
                                        {
                                                check if Ap_i′ holds in KB
                                                if successful
                                                {
                                                        if status of  D_i′ is
                                                        'observation' or 'assumed'
                                                        {
                                                                add current instantiation of
                                                                Arg_i to S
                                                        }
                                                        else
                                                        {
                                                                recursively get set S′ of all
                                                                arguments for  D_i′

                                                                if S′ is not empty
                                                                        attach S′ to D_i′
                                                                        and add current
                                                                        instantiation of
                                                                        Arg_i to S
}}}}}}}}
```

Figure 7. Pseudocode describing the first phase of argument generation.

Argument generation by interactive GenIE differs from that by the GenIE Assistant in two respects. First, interactive GenIE does not search for any arguments in support of the data of an argument until so requested by the user. Second, in interactive GenIE, an argument need not satisfy the applicability constraint of the argumentation scheme from which the argument was constructed. Instead, each condition of the applicability constraint is treated as a critical question that the user may ask to challenge the acceptability of the argument.

The user may request pro or con arguments, i.e. arguments respectively for or against a claim C. In response, the system returns the arguments including data D, warrant W, and set S of critical questions.[13] The user may then request pro or con arguments for any of these components,[14] and so on. Figure 9 shows a dialogue with Interactive GenIE (glossed in natural language for purposes of illustration) on a case of osteogenesis imperfecta (OI) that is believed by the healthcare provider to be more likely due to a new mutation (mosaicism) than to be an autosomal dominant inherited
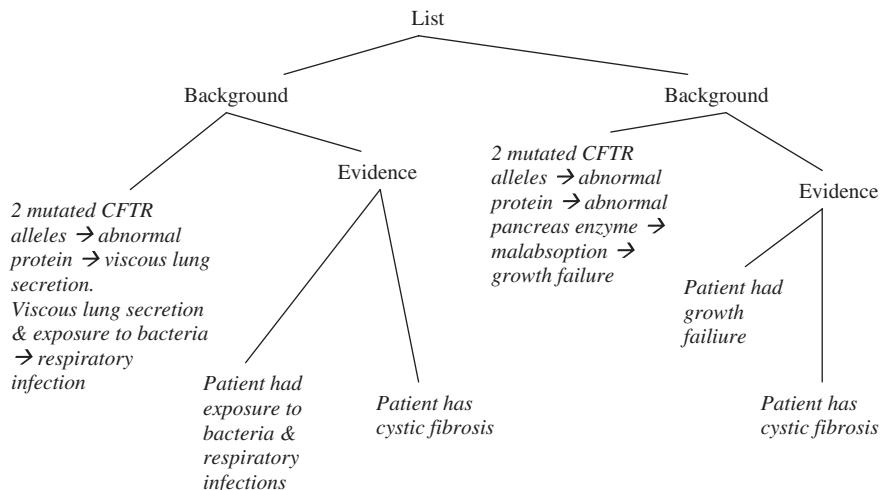
Figure 8. RST structure of arguments that the patient has CF. Left and right branches of background and evidence are satellite and nucleus, respectively.

1.	User:	Argument for OI mosaicism?
2.	GenIE:	*Claim:* OI mosaicism
3.		*Data:* Patient has flexible joints
4.		*Warrant:* Mutation associated with OI mosaicism causes the collagen protein produced by the affected cells to have an abnormal form leading to flexible joints
5.		*Critical question:* Could autosomal dominant OI be the cause of the flexible joints?
6.	User:	Argument for autosomal dominant OI?
7.	GenIE:	*Claim:* Not autosomal dominant OI
8.		*Data:* Test result for autosomal dominant OI was negative (normal amount of collagen protein)
9.		*Warrant:* Testing can show whether amount of collagen protein is low, and mutation associated with autosomal dominant OI causes a reduced amount of collagen protein to be produced

Figure 9. Example of dialogue with Interactive GenIE (glossed).

mutation.[15] Transparency of arguments is increased in Interactive GenIE compared with the GenIE Assistant by enabling the user to request arguments that may not be presented in a patient letter, and by making applicability constraints visible as critical questions that the user may pose.[16]

5.   Discourse grammar, argument presenter and linguistic realisation

This section covers the remaining components of the GenIE Assistant. Developing an application for a specific client was beyond the scope of the present work. These components were implemented to enable an evaluation of the generated arguments to be performed and to gain insight into challenges in presentation and realisation of transparent argumentation.

5.1.   *Pragmatic features of argument presentation*

First, we summarise the results of a study of pragmatic features of argument presentation in the corpus whose purpose was to inform the design of these components. Four letters in the corpus were

analysed with respect to three dimensions relevant to argument transparency: order of argument components, enthymemes[17] (arguments with implicit components), and discourse cues marking argument components. One striking observation is the prevalence of enthymemes in the letters, reducing the number of full arguments that could be studied.

In arguments containing both explicit data and claim, the data are always given before the claim, and if present the warrant is never given between the data and claim.[18] Warrants appear both before and after the data and/or claim. Subarguments (for the data) are given before the claim of the main argument.

As for when components are omitted, the data of an argument is not given explicitly when (1) it is given in a previous section of the letter, (2) it is in the common ground,[19] or (3) it is supported by (i.e. is the claim of) a subargument. Contrary to the view that warrants are often implicit, warrants are often explicit in these letters. Furthermore, the claim is often omitted. It is given explicitly when (1) it is a weaker claim (e.g. the preliminary diagnosis), or (2) when it is not likely to be negatively valued by the audience (e.g. the claim that neither parent is likely to carry the mutation responsible for their child's condition). Conveying negatively valued claims implicitly is consistent with the guideline in this genre to use indirection to mitigate the emotional impact of negatively valued information on the client (Baker et al. 2002), which we shall refer to later as the stylistic goal of conveying empathy.

Finally, there were few discourse cues marking data, claim, or warrant. In the four letters studied, there were two uses of cues ('therefore', 'then') preceding a claim, and one use of a cue ('since') preceding the data of an argument.

5.2. *Discourse grammar*

A discourse grammar was implemented based upon an RST analysis of several letters in the corpus and professional guidelines for writing patient letters (Baker et al. 2002). Since the goal of implementing the GenIE Assistant was to investigate argument generation, the discourse grammar was implemented to cover the topic sections in this genre that make the most extensive use of argumentation. The discourse grammar is a set of rules[20] (implemented in Prolog) that construct a partial DPlan representing the RST structure and content of a document, not including the arguments. Rules of the grammar generate trees of RST relations. Leaves of a tree consist of content, in non-linguistic form, extracted by discourse grammar rules from the KB. For example, Figure 10 shows the structure generated by the grammar for the first section of a patient letter. In the figure, event propositions are glossed in italics.

Several discourse grammar rules are shown in simplified form in Figure 11. Rule 1 describes the four main sections of a letter. Rule 2 describes the four subsections of the first section, and the structuring of these four subsections as elements of an RST Narration relation. Rule 3 describes the extraction of the patient's symptoms from the KB to create the left-most subtree in Figure 10, describing the reason for the patient's referral to the clinic.

For each leaf in the DPlan that would be the claim of an argument, the argument generator is invoked (by a function call from the right-hand side of a grammar rule). The resulting RST structure, a list of all arguments for that claim, is inserted into the RST structure of the document instead of just the claim. For instance in Figure 10, the shaded leaf shows where the argument for that claim, shown in Figure 8, would be added to the DPlan. In this way, genre-specific discourse planning, implemented by the discourse grammar, is distinguished from argument generation. However, the content surrounding the arguments as well as the arguments are represented uniformly in RST to facilitate argument presentation and linguistic realisation.
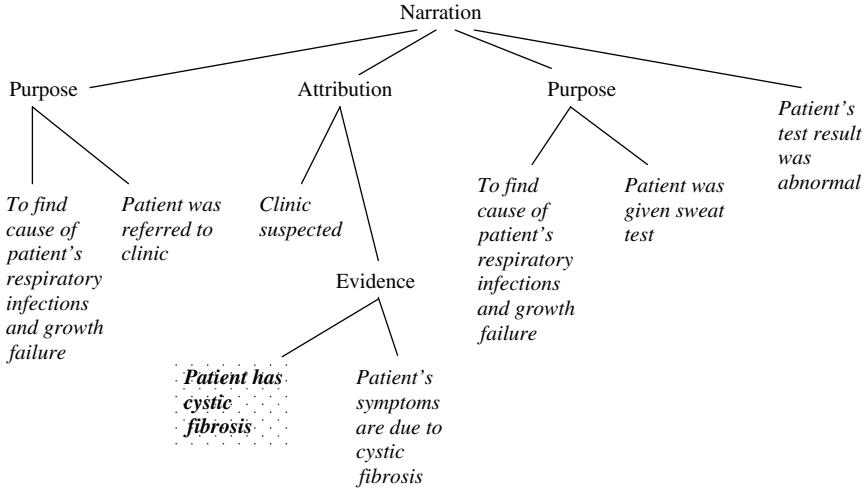
Figure 10. RST structure of first section of letter on patient who has CF. Left and right branches are satellite and nucleus respectively. Argument in Figure 8 replaces shaded node.

(1) letter(P,D,S,R) → pretest(P), diagnosis(D), source(S), risk(R).

(2) pretest(narration(T1, T2, T3, T4)) → referral(T1), pretest_diagnosis_act(T2),
                                          testing(T3), test_result(T4).

(3) referral(purpose(nuc(e1), sat(e2))) → get_symptoms(patient, Sym),
                                          % e1: patient was referred to clinic
                                          make_event(e1, …),
                                          % e2: to diagnose cause of Sym
                                          make_event(e2, …, Sym).

Figure 11. Some discourse grammar rules (simplified).

5.3. *Argument presenter*

In the GenIE Assistant's architecture, the argument generator's responsibility for generating an argument is distinguished from the argument presenter's responsibility for making decisions about how an argument is to be presented. More specifically, its responsibility is to aggregate or delete components of arguments in the DPlan for the sake of discourse coherence and argument transparency in the generated text.[21] Unfortunately, coherence and transparency goals may conflict. To maximise transparency, each argument's data, warrant, and claim would be given explicitly. However, a commonly used NLG strategy for improving coherence of generated text is to make it more concise through aggregation and deletion (Reiter and Dale 2000).

The purpose of the study described in Section 5.1 was to use the corpus to gain insight into presentation strategies including when to omit information. However, it was found that many of the arguments avoid making explicit claims of negatively valued information. That is, in the corpus, in addition to coherence goals, empathy may take precedence over transparency. Since empathy is not a stylistic goal for the GenIE Assistant, it is not clear when the corpus should be used to inform the decision to delete components of arguments from the DPlan. Therefore, the argument presenter was implemented using hand-crafted heuristic rules that were refined by trial and error. The remainder of this section describes the six main heuristics implemented in this version of the system. As will be seen in Section 6, however, additional heuristics are needed.

One heuristic aggregates two conjoined propositions in the claim of an argument when they share the same logical predicate-argument structure. For example, the claim glossed as: *the mother has one mutated copy of CFTR and the father has one mutated copy of CFTR* will be aggregated into a structure glossed as: *the mother and the father each has one mutated copy of CFTR*.

Another heuristic removes a claim *P* derived by the *conjunction simplification* argumentation scheme (Table 5) if it would immediately follow the data (*P* and *Q*) of that scheme. For example, if the data of an argument is the proposition glossed as *the mother has at least one mutated allele of CFTR and the father has at least one mutated allele of CFTR*, the claim glossed as *the mother has at least one mutated allele of CFTR* would be pruned since it repeats part of the data of the argument.

In one of the test cases generated by the GenIE Assistant, a letter on FH, three consecutive arguments for the claim *C* that the patient could have FH are given. The data given in the first, second, and third argument are, respectively, D_0 and D_1 (*The patient had a myocardial infarction and is a smoker*), D_0 and D_2 (*The patient had a myocardial infarction and is obese*), and D_0 and D_3 (*The patient had a myocardial infarction and has a low level of physical activity*). In other words, some of the data, D_0, were repeated in each argument. Thus, a pruning heuristic is used to avoid repeating D_0 in the second and third consecutive arguments for *C*. In addition, the warrants of the three arguments have some overlapping content, which is handled by the following heuristic.

This heuristic is used to avoid repeating part of the same warrant in two or more consecutive arguments for the same claim *C*. Suppose that the warrant of the first argument for *C* is that *G* can cause *P* and *P* can cause S_1 (e.g. *CF can cause abnormal CFTR protein. Abnormal CFTR protein can cause a viscous lung secretion.*), and the warrant of the second argument for *C* is that *G* can cause *P* and *P* can cause S_2 (e.g. *CF can cause abnormal CFTR protein. Abnormal CFTR protein can cause an abnormal pancreas enzyme level.*). In the consecutive arguments for *C* following the first argument, the duplicated first part of the warrant, *G* can cause *P*, is pruned and marked for realisation using an adverbial such as *also*. The beginning of the warrant of the second argument in the preceding example would then be realised as *Also, abnormal CFTR protein can cause an abnormal pancreas enzyme level* (e.g. see paragraph one in Figure 12).

A different heuristic is used to avoid repeating the same warrant in two consecutive, similar arguments referring to different individuals. Suppose that the warrant of each of the two arguments is that *G* can cause *S* (e.g. *One changed copy of FGFR3 can cause Achondroplasia symptoms*). Now suppose that the data and claim of the first argument are, respectively, D_1 (e.g. *The patient's mother does not have any symptoms*) and C_1 (e.g. *The patient's mother is not a carrier)* referring to individual I_1. Suppose that the data and claim of the second argument are, respectively, D_2 (e.g. *The patient's father does not have any symptoms*) and C_2 (e.g. *The patient's father is not a carrier*), where D_2 and C_2 are identical to D_1 and C_1, respectively, except that D_2 and C_2 refer to individual I_2. In the second argument, the warrant is pruned from the DPlan, and marked for realisation using an adverbial such as 'similarly'. In this example, the second argument would then be realised as *Similarly, the patient's father does not have any symptoms. Thus he is not a carrier* (e.g. see paragraph three in Figure 13).

While the above heuristic recognises two consecutive, analogous atomic arguments about two different individuals, in some cases, the DPlan contains two consecutive, analogous complex argument structures about two different individuals. A variation of the above heuristic is used in such cases to reduce the second argument structure to the 'similarly' adverbial, the data of the second argument, and the claim of the second argument (e.g. see paragraph three in Figure 12).

In summary, the above heuristics act upon elements of individual arguments or sets of arguments. In contrast, previous NLG research has focused on identifying semantic and/or syntactic structures that license aggregation (Shaw 1998; Dalianis 1999; Hielkema 2005; Harbusch and Kempen 2009).

The patient was referred to the clinic in order to diagnose the cause of his respiratory infections and growth failure. The clinic thought that respiratory infections and growth failure could be due to cystic fibrosis. Cystic fibrosis can cause abnormal CFTR protein. Abnormal CFTR protein can cause a viscous lung secretion. Exposure to bacteria and a viscous lung secretion can lead to respiratory infections. He had respiratory infections. He had exposure to bacteria. Thus he could have cystic fibrosis. Also abnormal CFTR protein can cause an abnormal pancreas enzyme level. An abnormal pancreas enzyme level can cause malabsorption. Malabsorption can cause growth failure. He had growth failure. Thus he could have cystic fibrosis. Therefore respiratory infections and growth failure could be due to cystic fibrosis. The clinic performed a sweat test in order to determine if he could have cystic fibrosis. The sweat test showed an abnormal NaCl level.

Cystic fibrosis is a disease caused by having two changed copies of a gene called CFTR. Two changed copies of the CFTR gene can cause abnormal CFTR protein. Abnormal CFTR protein can lead to an abnormal NaCl level and a sweat test can detect an abnormal NaCl level. The patient has an abnormal NaCl level. Thus he has cystic fibrosis. Therefore his respiratory infections and growth failure are due to his two changed copies of the CFTR gene.

North European ancestry increases the risk of having at least one changed copy of the CFTR gene. The patient's mother has north European ancestry. Thus she could have one changed copy of the CFTR gene. A child with two changed copies of a gene inherited one changed copy from the mother and one changed copy from the father. The patient has cystic fibrosis. Thus the patient's mother and the patient's father each have at least one changed copy of the CFTR gene. Two changed copies of the CFTR gene can cause cystic fibrosis symptoms. The patient's mother does not have cystic fibrosis symptoms. Thus she does not have two changed copies of the CFTR gene. Therefore she could have one changed copy of the CFTR gene. Similarly the patient's father has north European ancestry. He does not have cystic fibrosis symptoms. Thus he could have one changed copy of the CFTR gene.

Suppose the patient's mother has one changed copy of the CFTR gene and the patient's father has one changed copy of the CFTR gene. A twenty-five percent chance exists that future siblings will inherit two changed copies of the CFTR gene and they will have cystic fibrosis symptoms. Alternatively a seventy-five percent chance exists that they will not inherit two changed copies of the CFTR gene and they will not have cystic fibrosis symptoms.

Figure 12. Letter on CF case generated by GenIE Assistant.

5.4. *Linguistic realisation*

The linguistic realisation component transforms the DPlan, a forest of trees of RST relations and event propositions, into English text. An event proposition is described in terms of its modality and semantics, which is described by an action predicate and semantic case roles such as agent and beneficiary. The first phase of linguistic realisation maps the DPlan to a list of protosentences. A protosentence specifies one or more event propositions to be expressed in the same sentence in the final output. A protosentence that includes more than one event proposition specifies the rhetorical relation(s), e.g. Purpose, relating the events. The second phase of linguistic realisation maps each protosentence to a sentence specification. In this phase, the main NLG tasks performed are lexical choice and referring expression generation. The third phase of realisation maps the sentence specification to the input specification for SimpleNLG, an off-the-shelf surface realisation component (Gatt and Reiter 2009). Finally, SimpleNLG is used for morphological-syntactic processing.

In the rest of this section, only the aspects of linguistic realisation specific to argumentation will be described. During the first phase, clausal order of components of an argument is determined.

The patient was referred to the clinic in order to diagnose the cause of his disproportionately short arms and legs. The clinic thought that these symptoms could be due to achondroplasia. Achondroplasia can cause abnormal FGFR3 protein. Abnormal FGFR3 protein can cause abnormal bone development. Abnormal bone development can cause disproportionately short arms and legs. He had these symptoms. Thus he could have achondroplasia. The clinic performed a DNA test in order to determine if he could have achondroplasia. The DNA test showed a G1138A mutation.

Achondroplasia is a condition caused by having one changed copy of a gene called FGFR3. One changed copy of the FGFR3 gene can take the form of a G1138A mutation and a DNA test can detect a G1138A mutation. The patient has a G1138A mutation. Thus he has achondroplasia. Therefore his disproportionately short arms and legs are due to his one changed copy of the FGFR3 gene.

Being over 35 years old increases the risk of having a new change in a reproductive cell called a germline mutation. The patient's father is 45 years old. Thus he could have a germline mutation. A child with one changed copy of a gene inherited it from his mother or from his father. The mother or father may have inherited the change and have it in every cell, or the parent may have a germline mutation. The patient has achondroplasia. Thus, either the patient's mother or the patient's father has one changed copy of the FGFR3 gene in every cell or the patient's father could have a germline mutation. One changed copy of the FGFR3 gene can cause achondroplasia symptoms. The patient's mother does not have achondroplasia symptoms. Thus she does not have one changed copy of the FGFR3 gene in every cell. Similarly the patient's father does not have achondroplasia symptoms. Thus he does not have one changed copy of the FGFR3 gene in every cell. Therefore he could have a germline mutation.

Figure 13. Letter on achondroplasia case generated by GenIE Assistant.

Following the ordering observed in the corpus (as noted in Section 5.1), data (including subarguments) precede the claim. Although no pattern was observed in favour of providing the warrant before or after data and claim, it was decided to provide the warrant before data and claim since, as represented by its role of satellite in a Background relation, it provides background on the connection between data and claim.

In addition to determining clausal order, the first phase of realisation may add features to a protosentence for lexicalisation of discourse cues in the second phase of realisation. Informed by the use of such cues in the corpus, the only argument-specific information added is a feature subsequently realised as a sentence premodifier marking the claim of an argument. (As described in Section 5.3, features marking pruned components of arguments that are subsequently realised with adverbials 'also' and 'similarly' are determined by the argument presenter.) The feature is lexicalised as 'thus' for the claim of a subargument, and 'therefore' for a top-level argument. An example can be seen in the third paragraph of the generated letter shown in Figure 12. This paragraph contains three subarguments, each of whose claims is prefixed with 'thus', followed by the main claim, *Therefore she could have one changed copy of the CFTR gene*.

6. Evaluation of the GenIE Assistant

6.1. *Goals and design of evaluation*

A small controlled study of letters generated by the GenIE Assistant compared with letters written by a genetic counsellor on each of two medical cases was performed to determine if the system was generating argumentation 'about as good' as that written by the counsellor, and if not, to suggest areas for improvement in the Assistant. As a simple way of indirectly assessing argumentation, the

study determined the number of edits made by domain experts to letters produced by the GenIE Assistant compared with letters written by a genetic counsellor. The participants performing the editing task were university graduate students, at least 18 years old, majoring in biology-related fields, who had taken college-level genetics courses,[22] and who were screened for English writing fluency. Participants were compensated for their time.

The evaluation was performed using letters on two fictitious patients with substantially different conditions – CF and achondroplasia, which together cover all three mechanisms of inheritance of single-gene autosomal genetic disorders. Information for the cases was obtained from a genetics reference book (Nussbaum, McInnes, and Willard 2001). Seventeen participants were randomly assigned to one of four conditions representing the four combinations of authorship (genetic counsellor or GenIE Assistant) and genetic condition (CF or achondroplasia).[23]

It was decided to evaluate GenIE's argumentation with an editing task for a number of reasons. First, if we had not presented participants with the output of end-to-end generation, but instead had rendered the generated letters (or just the arguments) into English text ourselves, we might have inadvertently affected the judgment of the participants. Second, the editing task did not require the participants to have had training in argumentation. Third, the editing task has some ecological validity since, if the GenIE Assistant were deployed, it is expected that its output would be edited by a genetic counsellor to ensure accuracy as well as readability. Editing was used by Sripada, Reiter, and Hawizy (2005) to evaluate an operational NLG system that produces draft weather reports which are edited by human weather forecasters before being released to readers.[24] Finally, the results of the editing task could provide insight into areas of needed improvement, e.g. in argument presentation and linguistic realisation.

6.2. *Materials and procedure*

To generate a text for evaluation purposes, first, it is necessary to implement a KB[25] representing the relevant genetic conditions[26] and update it with the experts' presumed beliefs about the patient's case. No changes in the argument generator are required to handle different genetic conditions or patient cases.[27] The argument presenter's heuristics should be applicable to other genetic conditions. The linguistic realisation component was implemented to enable the formal evaluation to be performed; therefore parts of it, such as lexicalisation rules, have only been implemented to cover test cases.

Each of the two human-authored letters used in the study was created by asking a professional genetics counsellor to write a patient letter given the same findings, diagnosis, etc. given to the GenIE Assistant for each letter. To control for length, subtopics in the human-authored letters not covered in the three sections of GenIE's discourse grammar were removed by the experimenter. The two letters that were generated by GenIE[28] and used in the study are shown in Figures 12 and 13. The letter generated on CF contains all of the same types of arguments, except one,[29] as a letter in the corpus on an analogous case of a patient with an autosomal recessive disorder (hearing loss due to GJB2 mutation) inherited from parents who were carriers. The letter generated on achondroplasia contains exactly the same types of arguments with the same argument content as a letter on a similar case of achondroplasia in the corpus.

Each participant was given a copy of a letter, printed double-spaced on paper, and was asked to read it and manually edit the paper copy to make the information more understandable to people without a background in medicine or genetics. Participants were not informed that the letter may have been generated by computer.

Table 9. Mean number of edits.

CF	Reorder	Agg/disagg	Add	Delete	Reword	Total edits (SD)
GenIE	3.8	5	0.8	8.2	11.8	29.6 (9.39)
Human	1	2	2.75	0	9.75	15.5 (4.04)
Achondroplasia						
GenIE	4.25	4.25	1	3	8.25	20.75 (2.06)
Human	2.25	0.5	2	0	8.25	13 (4.08)

6.3. *Results and discussion*

To score the editing task, each participant's edits were classified and counted by the first author of this paper as follows:

- Deletion of a clause counts as one deletion (*Delete*).
- Addition of a clause counts as one addition (*Add*).
- Two clauses combined into one, e.g. through deletion and rewording, or one clause transformed into two clauses counts as one aggregation or disaggregation, respectively (*Agg/Disagg*).
- Rewording (change, addition, or deletion) of one or more contiguous words in a clause counts as one rewording (*Reword*).
- Moving one or more contiguous words in a clause counts as one reordering (*Reorder*).

All categories except reordering are mutually exclusive; e.g. a string of text can be counted as both rewording and reordering, but not as both rewording and aggregation. The experimenter's judgment as to the participant's intention was used to decide in cases where more than one classification might apply by the above criteria. Note that it would be quite challenging to assess argument content, presentation and realisation independently of each other and independently of the rest of the letter. The above scoring method conflates these factors. For example, the counted edits may include edits to text in the letter that is not part of any argument. Argument content and presentation is conflated since no distinction is made, e.g. between deletion of duplicate argument content and deletion of non-duplicate argument content. However, given the formative goals of the evaluation, it was decided to adopt a simple approach to scoring.

The 'bottom-line' results of the evaluation are not surprising, as shown in Table 9. For each letter, the mean total number of edits was higher for the GenIE-produced letter than its human-authored counterpart.[30] This is not surprising since the experiment compared the output of a human writer with the output of a proof-of-concept NLG system. However, it is important to look at the means for each category of edits to see what the experiment suggests about argument content as well as what areas of improvement might have the greatest impact on the results.

The categories that are most relevant to assessing GenIE's argument content are adds and deletes. In terms of adds, subjects actually made *fewer* additions to GenIE-produced letters than to human-authored letters. On review of the delete edits made to the GenIE-generated letters, it was observed that the majority were clauses that *repeated* data or claims given elsewhere in the letter. (No delete edits were made to the human-authored letters.) It would be straightforward to implement additional argument presenter heuristics to prune such clauses.

On review of the agg/disagg edits, it was observed that almost half involved combining two clauses by making one a relative clause of the other and that most did not involve argumentation. Those that did involve argumentation concerned combining arguments about the patient's mother and father, i.e. instead of discussing each of the parents individually. Finally, analysis of rewording

edits suggests that overall text quality could be substantially improved by modifications to the linguistic realisation of referring expressions.

The results suggest that, although improvement is needed to eliminate redundancy and polish the expression in natural language, the argument content produced by the Assistant is comparable to that produced by the genetic counsellor.

7. Related work

Although domain knowledge is represented in the GenIE Assistant with QPNs, the goal of our work differs from the research on explanation of reasoning in probabilistic expert systems, e.g. (Druzdzel 1996). The goal of that field is to make an expert system's reasoning plausible to human experts by describing the structure of its network and explaining its probability calculations (Lacave and Díez 2002). In contrast, the goal of our work is to generate arguments that are transparent to a lay audience. Thus, our computational model has been informed by analysis of a corpus of arguments written for a lay audience. Although the analysis revealed a causal conceptual model of the domain underlying expert–client communication, the arguments that were found are not concerned with explaining probability calculations. Instead, they are composed from presumptive argumentation schemes used in everyday communication as well as science and law (Walton et al. 2008).

Research on argumentation for genomic medicine applications includes risk assessment in genetics (Coulson, Glasspool, Fox, and Emery 2000), a rule-based decision-support tool to assist doctors in assessing the qualitative risk that a patient has a particular mutation, given data about the patient's family tree. The system's conclusion as to the patient's level of risk is explained to a doctor by listing evidence for and against the assessment. While risk assessment is outside of the scope of our research, the arguments that can be given by the GenIE Assistant are more varied and more complex. The REACT system (Glasspool, Fox, Oettinger, and Smith-Spark 2006) is a decision-support system for planning the medical care of women diagnosed as at risk of developing cancer due to genetic factors. REACT presents arguments for and against treatment options. However, the arguments are not generated 'on the fly' but are human-authored.

Although our work has been informed by other NLG research on argument generation, previous research has focused on the goal of persuasion or evaluation. Early work on persuasion addressed linguistic realisation (Elhadad 1992). Some NLG research has employed probabilistic network formalisms to compute persuasiveness (Zukerman, McConachy, and Korb 1998, 2000; Zukerman, McConachy, Korb, and Pickett 1999; Carofiglio and de Rosis 2003; Mazzotta and de Rosis 2006). In work on generating evaluative arguments, Carenini and Moore (2006) use an approach based upon the decision theory to select evidence. While the GenIE Assistant has a qualitative probabilistic domain model, it is not used to compute persuasiveness or utility.

Other NLG research on generation of persuasive arguments has used non-probabilistic approaches. Reiter, Robertson, and Osman (2003) used genre-specific discourse strategies. Grasso, Cawsey, and Jones (2002) dialogue system plans arguments using strategies from the New Rhetoric (Perelman and Olbrechts-Tyteca, 1969). Reed and Long's (1997, 1998) system operationalised standard deductive argument patterns as plan operators. Branting, Callaway, Mott, and Lester's (1999) model for drafting legal arguments used legal reasoning and genre-specific rhetorical strategies. Unlike these approaches, the approach presented in this paper uses non-genre-specific, non-domain-specific, mainly causal, normative argumentation schemes.

Some NLG research has addressed argument presentation issues. In Reed and Long's (1997) approach, ordering of argument components is determined both by planning goals and by heuristics derived from corpora and a survey of the literature on persuasion. Carenini and Moore (2006) empirically studied the optimal amount of evidence and its optimal ordering in designing evaluative

argument strategies. The approach to presentation described in this paper differs from these since the goal is not persuasion or evaluation but transparency. Since transparency may be adversely affected by verbosity, GenIE Assistant's pruning heuristics are motivated by the same goals as Fiedler and Horacek's (2001) model, which uses summarisation and omission of inferable details to make the presentation of machine-generated mathematical proofs more comprehensible.

8. Limitations and future work

The goal of our research was to investigate a computational model of transparent expert-to-lay argumentation rather than to develop a system meeting the requirements of a specific healthcare provider. To apply the model to build real-world systems, software engineering work remains including additional requirements elicitation (e.g. study of a corpus of the client organisation's patient letters), constructing KBs for the genetic disorders of interest, modifying the discourse grammar to handle all possible letter variants and stylistic preferences of the client organisation, and building more robust argument presentation and linguistic realisation components. Advances in NLG technology such as in aggregation and referring expression generation can be leveraged to improve expression of the letters. Furthermore, it may be advisable to incorporate NLG techniques developed for low-literacy audiences (Williams and Reiter 1998) into this type of application.

In terms of theoretical limitations, the current argument generator does not address the generation of information added to a patient letter to address possible misconceptions of a lay audience. For example, in one letter in the corpus, the writer points out that the patient's negative test result does not rule out the possibility that the patient's medical problems are due to some other genetic condition. A user model representing typical lay misconceptions about the domain is a prerequisite for handling this type of argument (Green 2005b). In addition, affective user modelling is a prerequisite for use of presentation strategies designed to mitigate the impact of negatively valued information (Green 2005b, 2005c). Work is in progress to add user modelling to the architecture. In addition, work is in progress to improve argumentation presentation (aggregation and pruning), to add argumentation schemes that provide backing for warrants, and to add graphics to the textual presentation of risk information as illustrated in Green, Britt, Jirak, Waizenegger, and Xin (2005).

Beyond genetic counselling, the approach presented in this article to argument generation (in print or in an interactive system) may be applicable to lay communication in other biomedical or scientific fields whose content can be represented in a qualitative causal KB. Furthermore, as automatic methods are developed for transforming Bayesian networks into QPNs (Druzdzel and Henrion 1993), it may be possible to adapt pre-existing Bayesian KBs for this purpose. Another possible new direction is to apply the argumentation schemes developed for the current project to the interpretation of causal argumentation in text.

9. Summary

This article describes a software architecture for NLG of biomedical argumentation for a lay audience. The architecture has been implemented and evaluated in the GenIE Assistant, a prototype system for drafting genetic counselling patient letters. A significant challenge in this genre is the range of domain knowledge, i.e. there are over 4500 single-gene autosomal genetic disorders. This challenge has been met by identifying an abstract causal domain model used in the field for communication with a lay audience. Use of the abstract model constrains design of KBs, simplifies content selection in argument generation, and enables analysis of argumentation in the corpus to abstract away from details of particular genetic conditions.

Another challenge is that although the conclusion of a normative argument may be defeasible, it must be supported by data and, in most cases, be warranted by scientific principles. In addition, a normative argument must meet other criteria that can be formulated as applicability constraints (i.e. critical questions). To support normative argument generation, abstract argumentation schemes were defined. They refine the causal argumentation schemes described by argumentation theorists (Walton et al. 2008) in terms of abstract properties of a causal KB and can be reused for argument generation in other causal domains.

Another challenge is the goal of transparency. Transparency is necessary before an audience can fully comprehend, evaluate, or challenge an argument, or re-evaluate it in light of new findings or changes in scientific knowledge. Transparency is supported by reconstructing the full argument for a claim using argumentation schemes that differentiate the roles of data, warrant, and critical questions of an argument. The implementation of the Interactive GenIE prototype labels the functional roles of the parts of an argument and enables the user to request pro and con arguments for its components. Generation of transparent arguments in a non-interactive medium is challenging since transparency may conflict with other NLG goals. As a first step, the GenIE Assistant's argument presenter makes use of its knowledge of argument structure in applying discourse aggregation and pruning heuristics. In addition, knowledge of argument structure was used in its linguistic realisation component to order the clauses in an argument and to mark the claims of arguments with discourse cue words.

Acknowledgements

Notes

1. We consider these patterns to represent argumentation in the sense that they are used to support a position that might be challenged by the audience at some time. The term 'audience' includes the primary audience, the client(s) to whom the letter is addressed, as well as persons reviewing the letter in case of questions as to the quality of service provided to the patient (Baker et al. 2002). For example, without transparent biomedical argumentation, parents of a child might be reluctant to accept the clinic's conclusion that their child's medical condition is due to a mutation that he inherited from them. For another example, a client may question an expert's conclusion at some future time when the biomedical principle that had been used to justify the conclusion has been rejected by new scientific studies. For a final example, argumentation for the preliminary diagnosis could be used later to justify what tests were performed in case the clinic is accused of providing inadequate care to the patient.
2. The GenIE Assistant is not being developed for deployment by a specific organisation. The goal of implementing the Assistant is to ground the research in a potential, significant, complex real-world application. It is assumed that if the GenIE Assistant were deployed, a healthcare provider could provide patient-specific evidence (symptoms, test results, etc.) and conclusions (e.g. the diagnosis) through menus and checklists in a graphical user interface such as the prototype shown in Green et al. (2005).

The GenIE Assistant would then use that information and general information from its KBs on different genetic disorders to draft the patient letter, including arguments for the given conclusions. It is assumed that the provider would review and (if necessary) edit the draft. However, the GenIE Assistant is not designed to automate medical reasoning tasks such as diagnosis, nor is it intended to replace the client's face-to-face meeting with a genetic counsellor.

3. The arguments are normative in the sense that, although they may have defeasible conclusions, they meet professional standards in the field of clinical genetics. In other words, they represent the arguments that a domain expert should provide to a lay person. This article does not address generation of non-normative arguments found in the corpus, whose purpose is to mitigate the emotional impact of negatively valued information on the audience (Green 2005b).

4. There are more than 4500 single-gene autosomal disorders, affecting about 1% of the population (Wilson 2000). Although the corpus informing this work may be small compared with corpora needed for machine learning, it was sufficient for our goals due to constraints of the genre and domain. The nine texts are representative of the genre as described in Baker et al. (2002) and contain a variety of argument patterns and complex compositions of argument patterns. As for the domain, the nine letters cover the three main mechanisms of inheritance of single-gene autosomal genetic disorders: autosomal recessive inheritance, autosomal dominant inheritance, and new mutations. (The rest of the corpus includes letters on chromosomal and multifactorial genetic disorders.)

5. See Green (2005a) for suggestions on how the process might be automated. However, there was no need to automate it for the objectives of this project. The purpose of modelling the content of letters, at first, was to study its role in argumentation schemes. Later, as described in Section 3, the set of eight concepts and their inter-relations were used to constrain design of KBs used in the GenIE Assistant.

6. It is not claimed that the grammar is sufficient to generate all possible genetic counselling patient letters. The purpose of implementing the grammar was to enable the implementation of the prototype system and the evaluation of generated arguments.

7. Including hearing loss due to GJB2 mutation, velocardiofacial syndrome, neurofibromatosis, osteogenesis imperfecta, phenylketonuria, familial hypercholesterolemia, achondroplasia, and cystic fibrosis (CF). Information on CF, a condition not covered in the corpus, was obtained from textbooks (Nussbaum et al. 2001). In addition, the textbooks were consulted to help interpret letters in the corpus.

8. Since it was beyond the scope of this project to deploy the GenIE Assistant, we did not develop tools for automated KB creation.

9. These relations are defined formally in Wellman (1990) and Druzdzel and Henrion (1993).

10. The distinction in GenIE between data and warrant corresponds, respectively, to the distinction between knowledge about a particular patient's case and the general biomedical knowledge linking the data to a claim. Although this distinction may not matter for some purposes, it is important at least from a software engineering perspective. From that perspective, the knowledge used for warrants is obtained by knowledge engineers, reusable, and to be revised by experts in response to changing beliefs in the field of biomedicine. In contrast, the knowledge used for data is patient-case-specific and subject to privacy restrictions, much more volatile, and provided by clinical staff.

11. To clarify how the example shown with this scheme, *P has 2 mutated GJB2 alleles*, instantiates the pattern of its claim, $A \geq a$, note that the pattern describes a pattern in the KB, not a textual pattern in the corpus. The pattern is that a KB variable has reached its threshold value. (The notation would be simpler if the domains of all variables were Boolean; however, in this field of knowledge, it is necessary to allow for other variable types.) In the example, the variable represents the number of mutated alleles in the patient's GJB2 genotype, where the threshold value is 2 (since it takes two mutated alleles to result in hearing loss).

12. The term 'applicability constraint' was chosen for consistency with standard usage in the natural language planning/generation literature since argumentation schemes are used in the GenIE Assistant for argument generation.

13. The moves are similar to those of the Toulmin Dialog Game (DG) (Bench-Capon, Geldard, and Leng 2000). However, unlike TDG, Interactive GenIE also allows the user to pose critical questions. Another difference is that Interactive GenIE is not designed to model the effect of a dialogue on the participants' commitments.

14. Since D or W may be a conjunction of premises and S is a set of critical questions, it is more accurate to say that the user may request a pro or con argument for any atomic element of these components.
15. Based upon a case in the corpus.
16. Interactive GenIE was originally implemented to perform component testing of the GenIE Assistant's argument generator and KBs. However, its virtues in terms of transparency and user control became apparent, and we now envision its use in interactive patient education systems or intelligent environments for learning about scientific argumentation.
17. The term 'enthymeme' has a long history in the logic and argumentation theory (Walton et al. 2008).
18. The fact that the warrant never comes between data and claim is consistent with the representation of the warrant as the satellite of a background relation whose nucleus is an Evidence relation describing data and claim, and the principle that branches of an RST analysis should not cross.
19. The term 'common ground' comes from the field of discourse studies and refers to information shared by the so-called speaker and hearer (i.e. the participants in a dialogue, or a writer and intended audience). For example, based upon one of the letters in the corpus, it is plausible to assume that it was in the common ground of the parents who brought their deaf child to the genetics clinic for diagnosis and the clinical staff who examined him that the patient is a child with hearing loss.
20. There are 48 Prolog rules in the current implementation, covering four sections of a letter. Section 1 describes the reason for referral, the preliminary diagnosis, what tests were performed, and the test results. Section 2 describes the final diagnosis. Section 3 describes the probable source of the patient's genetic condition, e.g. from whom it was inherited. Section 4 describes inheritance risks for other family members. The rules are essentially NLG schemas (Reiter and Dale 2000).
21. In this implementation of the GenIE Assistant, other presentation decisions (ordering of argument components and addition of some discourse cues) are made in the linguistic realisation module.
22. The graduate coursework in genetics should have provided sufficient background for the editing task.
23. Five participants received the GenIE-authored letter on CF; the other three groups had four participants each.
24. Sripada et al. (2005) note that, while providing metrics for quantifying the practical usefulness of a system, the number of edits does not always indicate problems in the generated text. For example, some edits may be the result of an individual forecaster's preferences, or may be required because of edits made to the preceding text. Our impression is that similar factors may have affected the data collected in our evaluation.
25. The KB used for the CF case, for example, had 26 nodes.
26. A KB could be reused for different patients of course.
27. Argument generation has been tested informally on several other genetic conditions.
28. The total time to generate the two letters (on a standard personal computer), a total of 780 words, is about 22 s.
29. A type of argument found in the corpus but not implemented in the version of GenIE described in this paper provides a probability statement (e.g. 'about 50% of children with severe to profound recessively inherited non-syndromic genetic hearing loss have a change in . . . GJB2') as backing (Toulmin 1998) for an implicit warrant, instead of the warrant (e.g. having two mutated GJB2 alleles can lead to hearing loss); this type of argument has been implemented in a subsequent version of the system.
30. The data was analysed with SPSS using a two-sided t-test with a 95% confidence interval for the null hypothesis that the mean number of edits would be the same for the GenIE-produced and the human-authored version of each letter. The difference in means was found to be statistically significant, with P-values of 0.028 (CF letter) and 0.015 (achondroplasia letter).

References

Ancker, J.S., and Kaufman, D. (2007), 'Rethinking Health Numeracy: A Multidisciplinary Literature Review', *Journal of the American Medical Informatics Association*, 14(6), 713–721.

Baker, D.L., Eash, T., Schuette, J.L., and Uhlmann, W.R. (2002), 'Guidelines for Writing Letters to Patients', *Journal of Genetic Counseling*, 11(5), 399–418.

Bench-Capon, T.J.M., Geldard, T., and Leng, P.H. (2000), 'A Method for the Computational Modelling of Dialectical Argument with Dialogue Games', *Artificial Intelligence in Law*, 8, 233–254.

Branting, L.K., Callaway, C.B., Mott, B.W., and Lester, J.C. (1999), 'Integrating Discourse and Domain Knowledge for Document Drafting', in *Proceedings of International Conference on Artificial Intelligence and Law*, pp. 214–220.

Carenini, G., and Moore, J.D. (2006). 'Generating and Evaluating Evaluative Arguments', *Artificial Intelligence*, 170, 925–952.

Carofiglio, V., and de Rosis, F. (2003), 'Combining Logical with Emotional Reasoning in Natural Language Argumentation', in *Assessing and Adapting to User Attitudes and Effect: Why, When and How?, Workshop Proceedings, 9th International Conference on User Modeling (UM'03)*, eds. C. Conati, E. Hudlicka, and C. Lisetti, Johnstown, PA, pp. 9–15.

Coulson, A.S., Glasspool, D.W., Fox, J., and Emery, J. (2000), 'Computerized Genetic Risk Assessment and Decision Support in Primary Care', *Informatics: The Journal of Informatics in Primary Care*.

Dalianis, H. (1999), 'Aggregation in Natural Language Generation', *Computational Intelligence*, 15.

Druzdzel, M.J. (1996), 'Qualitative Verbal Explanations in Bayesian Belief Networks', *Artificial Intelligence and Simulation of Behavior Quarterly*, 94, 43–54.

Druzdzel, M.J., and Henrion, M. (1993), 'Efficient Reasoning in Qualitative Probabilistic Networks', in *Proceedings of the 11th National Conference on AI (AAAI-93)*, pp. 548–553.

Elhadad, M. (1992), 'Generating Coherent Argumentative Paragraphs', in *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, pp. 638–644.

Fiedler, A., and Horacek, H. (2001), 'Argumentation in Explanations to Logical Problems', in *Proceedings of ICCS 2001*, Springer LNCS 2073, pp. 969–978.

Fox, J., Glasspool, D., Grecu, D., Modgil, S., South, M., and Patkar, V. (2007), 'Argumentation-based Inference and Decision-making – a Medical Perspective', *IEEE Intelligent Systems*, 22(6), 34–41.

Gatt, A., and Reiter, E. (2009), 'SimpleNLG: A Realisation Engine for Practical Applications', in *Proceedings of the 12th European Workshop on Natural Language Generation*, Athens, Greece, 30–31 March 2009, pp. 90–93.

Glasspool, D.W., Fox, J., Oettinger, A., and Smith-Spark, J. (2006), 'Argumentation in Decision Support for Medical Care Planning for Patients and Clinicians', in *Argumentation for Consumers of Healthcare: Papers from the AAAI Spring Symposium*, Menlo Park, CA: AAAI Press, pp. 58–63.

Gordon, T.F., Prakken, H., and Walton, D.N. (2007), 'The Carneades Model of Argument and Burden of Proof', *Artificial Intelligence*, 171(4), 875–896.

Grasso, F., Cawsey, A., and Jones, R. (2002), 'Dialectical Argumentation to Solve Conflicts in Advice Giving: A Case Study in the Promotion of Healthy Nutrition', *International Journal of Human-Computer Studies*, 53, 1077–1115.

Green, N. (2005a), 'A Bayesian Network Coding Scheme for Annotating Biomedical Information Presented to Genetic Counseling Clients', *Journal of Biomedical Informatics*, 38, 130–144.

Green, N. (2005b), 'Affective Factors in Generation of Tailored Genomic Information', in *User Modeling 2005 Workshop on Adapting the Interaction Style to Affective Factors*, Edinburgh.

Green, N. (2005c), 'Analysis of Linguistic Features Associated with Point of View for Generating Stylistically Appropriate Text', in *Computing Attitude and Affect in Text: Theory and Applications*, eds. J.G. Shanahan, Y. Qu, and J. Wiebe, Springer, pp. 33–40.

Green, N. (2007), 'A Study of Argumentation in a Causal Probabilistic Humanistic Domain: Genetic Counseling', *International Journal of Intelligent Systems*, 22(1), 71–93.

Green, N. (2008), 'Dialectical Argumentation in Causal Domains', in *Proceedings of 8th International Workshop on Computational Models of Natural Argument*, pp. 31–38.

Green, N.L. (2010), 'Representation of Argumentation in Text with Rhetorical Structure Theory', *Argumentation*, 24(2), 181–196.

Green, N., Britt, T., Jirak, K., Waizenegger, D., and Xin, X. (2005), 'User Modeling for Tailored Genomic e-Health Information', *User Modeling 2005 Workshop on Personalisation for eHealth*, pp. 31–40.

Harbusch, K., and Kempen, G. (2009), 'Generating Clausal Coordinate Ellipsis Multilingually: A Uniform Approach Based Upon Postediting', *Proc. of ENLG 09*.

Hielkema, F. (2005), *Performing Syntactic Aggregation using Discourse Structures*, Unpublished Master's thesis, University of Groniingen.

Korb, K.B., and Nicholson, A.E. (2004), *Bayesian Artificial Intelligence*, Boca Raton, FL: Chapman & Hall/CRC.

Lacave, C., and Díez, F.J. (2002), 'A Review of Explanation Methods for Bayesian Networks', *The Knowledge Engineering Review*, 17(2), 107–127.

Lindgren, H., and Eklund, P. (2005), 'Differential Diagnosis of Dementia in an Argumentation Framework', *Journal of Intelligent and Fuzzy Systems*, 16, 1–8.

Mann, W.C., and Thompson, S.A. (1988), 'Rhetorical Structure Theory: Toward a Functional Theory of Text Organization', *Text*, 8, 243–281.

Mazzotta, I., and de Rosis, F. (2006), 'Artifices for Persuading to Improve Eating Habits', in *Argumentation for Consumers of Healthcare: Papers from the AAAI Spring Symposium*, Menlo Park, CA: AAAI Press, pp. 76–85.

McLeod, K., and Burger, A. (2008), 'Towards the Use of Argumentation in Bioinformatics: A Gene Expression Case Study', *Bioinformatics*, 24, i304–i312.

Nielsen-Bohlman, L., Panzer, A.M., Hamlin, B., and Kindig, D.A. (2004), *Institute of Medicine. Health Literacy: A Prescription to End Confusion. Committee on Health Literacy, Board on Neuroscience and Behavioral Health*, Washington, DC: National Academies Press.

Nussbaum, R.L., McInnes, R.R., and Willard, H.F. (2001), *Thompson and Thompson Genetics in Medicine*, Philadelphia, PA: W. B. Saunders Company.

Perelman, C., and Olbrechts-Tyteca, L. (1969), *The New Rhetoric: A Treatise on Argumentation*, Notre Dame, IN: Notre Dame Press.

Rahati, A., and Kabanza, F. (2009), 'Persuasive Argumentation in a Medical Diagnosis Tutoring System', in *Computational Models of Natural Argument IX*, July 13, Pasadena CA, pp. 39–48.

Reed, C.A., and Long, D.P. (1997), 'Content Ordering in the Generation of Persuasive Discourse', in *Proceedings of 15th International Joint Conference on Artificial Intelligence*, Nagoya, Japan, pp. 1022–1027.

Reed, C.A., and Long, D.P. (1998), 'Generating the Structure of Argument', in *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL98)*, ed. W. Hoeppner, Montreal, Canada, pp.1091–1097.

Reiter, E., and Dale, R. (2000), *Building Natural Language Generation Systems*, Cambridge: Cambridge University Press.

Reiter, E., Robertson, R., and Osman, L. (2003), 'Lessons from a Failure: Generating Tailored Smoking Cessation Letters', *Artificial Intelligence*, 144, 41–58.

Shaw, S. (1998), 'Segregatory Coordination and Ellipsis in Text Generation', *Proc. of 17th COLING*, Montreal.

Sripada, S.G., Reiter, E., and Hawizy, L. (2005), 'Evaluation of an NLG System using Post-edit Data: Lessons Learnt', in *Proc. of ENLG-2005*.

Toulmin, S.E. (1998), *The Uses of Argument* (9th ed.), Cambridge: Cambridge University Press.

Walton, D., Reed, C., and Macagno, F. (2008), *Argumentation Schemes*, Cambridge: Cambridge University Press.

Wellman, M.P. (1990), 'Fundamental Concepts of Qualitative Probabilistic Networks', *Artificial Intelligence*, 44(3), 257–303.

Williams, S., and Reiter, E. (1998), 'Generating Basic Skills Reports for Low-skilled Readers', *Natural Language Engineering*, 1(1), 1–32.

Zukerman, I., McConachy, R., and Korb, K. (1998), 'Bayesian Reasoning in an Abductive Mechanism for Argument Generation and Analysis', in *Proceedings of the 15th National Conference on Artificial Intelligence*, Madison, Wisconsin, pp 833–838.

Zukerman, I., McConachy, R., and Korb, K. (2000), 'Using Argumentation Strategies in Automated Argument Generation', in *Proceedings of the 1*st *International Natural Language Generation Conference*, pp. 55–62.

Zukerman, I., McConachy, R., Korb, K., and Pickett, D. (1999), 'Exploratory Interaction with a Bayesian Argumentation System', in *Proceedings of the 16*th *International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 1294–1299.