‘Healthy’ Coreference: Applying Coreference Resolution to the Health Education Domain

by

David Z. Hirtle

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2008

Abstract

This thesis investigates coreference and its resolution within the domain of health education. *Coreference* is the relationship between two linguistic expressions that refer to the same real-world entity, and resolution involves identifying this relationship among sets of referring expressions. The coreference resolution task is considered among the most difficult of problems in Artificial Intelligence; in some cases, resolution is impossible even for humans. For example, *she* in the sentence *Lynn called Jennifer while she was on vacation* is genuinely ambiguous: the vacationer could be either Lynn or Jennifer.

There are three primary motivations for this thesis. The first is that health education has never before been studied in this context. So far, the vast majority of coreference research has focused on news. Secondly, achieving domain-independent resolution is unlikely without understanding the extent to which coreference varies across different genres. Finally, coreference pervades language and is an essential part of coherent discourse. Its effective use is a key component of easy-to-understand health education materials, where readability is paramount.

No suitable corpus of health education materials existed, so our first step was to create one. The comprehensive analysis of this corpus, which required manual annotation of coreference, confirmed our hypothesis that the coreference used in health education differs substantially from that in previously studied domains. This analysis was then used to shape the design of a knowledge-lean algorithm for resolving coreference. This algorithm performed surprisingly well on this corpus, e.g., successfully resolving over 85% of all pronouns when evaluated on unseen data.

Despite the importance of coreferentially annotated corpora, only a handful are known to exist, likely because of the difficulty and cost of reliably annotating coreference. The paucity of genres represented in these existing annotated corpora creates an implicit bias in domain-independent coreference resolution. In an effort to address these issues, we plan to make our health education corpus available to the wider research community, hopefully encouraging a broader focus in the future.

Acknowledgements

I gratefully acknowledge the financial support of NSERC and the University of Waterloo, without which this thesis would never have been started let alone finished.

Special thanks to my thesis supervisor, Chrysanne DiMarco, for accommodating my interests (unfocused as they were at times) and, in fact, motivating me to come to Waterloo in the first place. I'm sure that I speak on behalf of many in saying that her continued efforts at sustaining computational linguistics at UW, through graduate courses and research projects alike, are greatly appreciated.

A number of others also deserve my thanks for being kind enough to share their experience, time and resources with me:

- My readers, Otman Basir and Randy Harris
- Olga Gladkov for her linguistic expertise
- Andy Chiu for answering my machine learning questions
- Constantin Orăsan (University of Wolverhampton) for providing me with the newest version of his annotation tool

I also want to extend my appreciation to family and friends for their support, and for inspiring many of the examples found throughout this thesis. (Yes, your name is probably in here somewhere.)

Last but certainly not least, I am deeply indebted to my loving wife Lacey for her endless patience, timely encouragement and continual assistance over the course of this work, not to mention the rest of the time. Ευχαριστώ!

Je n'ai fait celle-ci plus longue que parce que
je n'ai pas eu le loisir de la faire plus courte.[†]

Blaise Pascal

[†] "I have only made this long because I did not have the opportunity to make it shorter."

Table of Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

Coreference resolution is regarded as one of the most difficult problems in Artificial Intelligence. Consider the following examples:

(1.1) *Janet enjoys long walks with her husband.*

(1.2) *Acute viral nasopharyngitis is also known as the common cold.*

(1.3) *Jim called David while he was at the hospital.*

In the first case, *Janet* and *her* refer to the same person named 'Janet'. In other words, these two expressions **corefer** or have the relationship of **coreference**. Determining that *Janet* and *her* corefer is known as **coreference resolution**.

The second example illustrates a more complex instance of coreference between two noun phrases: *Acute viral nasopharyngitis* and *the common cold*. Resolution in this case involves knowing that these expressions are synonymous, referring to the same disease.

In the final example (1.3), coreference resolution is even more challenging because it is unclear who is at the hospital: the pronoun *he* could be referring to either Jim or David. The ambiguity of the pronoun makes either interpretation perfectly legitimate; there is no definitive answer without additional context.

The point is that coreference resolution can be tricky and sometimes, in the worst case, impossible—even for humans!

There has been, and continues to be, a great deal of interest and work on this problem. Coreference has received so much attention because it is an essential part of coherent text. For example, it takes longer to read stories with repeated proper names instead of appropriate pronouns [61]. Coreference resolution has applications in practically every area of natural language processing, including (but certainly not limited to) the following:

Information Extraction: Important data within a document can be 'hidden' by referring expressions, making it difficult to automatically identify and extract.

Text Summarization: Any sentence chosen to form part of the summary may be referentially dependent on nearby sentences, leading to disfluencies unless resolved correctly.

Natural Language Generation: Producing coherent discourse requires appropriate use of cohesive devices, including pronouns.

This thesis investigates coreference and its resolution specifically within the domain of health education. This focus is motivated by the following three insights:

- Health education has never before been studied in the context of coreference resolution. So far, the vast majority of research has been restricted to news articles and, to a lesser extent, technical manuals.
- The domain of health education has its own unique characteristics that distinguish it from previously studied domains. Achieving domain-independent coreference resolution is unlikely without understanding the extent to which coreference varies across different genres of text.
- Readability is paramount in health education, and the effective use of coreference is an important element of easy-to-understand text.

We set the stage by first introducing and exemplifying relevant concepts then reviewing related work in Chapter 2. The following three chapters of the thesis each detail a particular way in which this work contributes to existing research on coreference resolution. Specifically, these contributions are as follows:

1. **Creation of a corpus representing health education** (Chapter 3)

   There are no suitable corpora of health education texts, so we create our own by downloading materials on a wide range of topics from reliable sources such as Health Canada. We manually add coreference annotations to facilitate later analysis.

2. **Analysis of a new domain** (Chapter 4)

   We perform a comprehensive analysis of the annotated health education corpus, gathering and sifting through relevant data to reveal important characteristics of this domain. This information later proves useful during algorithm design.

3. **Coreference resolution algorithm** (Chapter 5)

   We develop an algorithm for resolving coreference among general noun phrases and evaluate its performance on our corpus. The results of these tests reveal even more about the health education domain.

The thesis concludes with a summary of accomplishments and possibilities for future work in Chapter 6.

Our hypothesis is that the coreference found in health education is less complex than that found in previously studied domains and therefore better-suited to automatic resolution. Our experiments will attempt to reveal whether or not this is the case.

Chapter 2

Background

Before the bulk of this thesis can be profitably discussed, a few preliminary details must be covered. First, we introduce key concepts and terminology related to coreference. With this foundation in place, we overview common strategies for handling coreference in Section 2.2. Finally, related annotation and resolution research is discussed in Section 2.3.

2.1 Terminology

The phenomenon of coreference has been studied from many different perspectives, including those of psychology, philosophy and (computational) linguistics. It should therefore not be surprising that terminology is often used inconsistently. In an attempt to avoid any such confusion in this thesis, we first define coreference and related concepts.

2.1.1 Reference

Any linguistic expression that refers to something (e.g., a person, organization, object or concept) in the real world is called a **referring expression**, and whatever is being referred to is called its **referent**. Often referents are entities, but they can also be actions, situations and events as in the following examples[1]:

(2.1) *See your doctor if you are feeling sick or worried, or if the test instructions recommend* ***doing so****.*

[1] From "Medical Test Kits for Home Use", "Whooping Cough (Pertussis)" and "Asthma"[17], respectively.

(2.2) *Older members of a household may have whooping cough without even realizing **it**.*

(2.3) *During an attack, the muscles around the airways can tighten and the airways can produce mucus. **These conditions** make it even harder to breathe.*

As with most existing research, this thesis focuses on reference involving noun phrases, i.e., referring expressions that are noun phrases and referents that can be represented by noun phrases. A **noun phrase** (NP) is a word or group of words that acts as a noun in a sentence and that revolves around a single noun or pronoun, which is called the **head** of the phrase. Passage (2.3), for instance, contains a total of six noun phrases (heads in bold):

NP_1 = *an **attack***

NP_2 = *the **muscles** around the airways*

NP_3 = *the **airways***

NP_4 = ***mucus***

NP_5 = *These **conditions***

NP_6 = ***it***

Each of these NPs falls into one of the following four categories:

- **Indefinite noun phrase**: an NP that begins with the indefinite articles *a* or *an* (e.g., NP_1) or a few other kinds of words including quantifiers (e.g., *most diets*) and cardinal numbers (e.g., *three fingers*). Indefinite NPs normally do not involve specific reference.
- **Definite noun phrase**: an NP that begins with definite article *the* (e.g., NP_2, NP_3), a demonstrative pronoun (e.g., NP_5), a possessive pronoun (e.g., *your condition*) or a proper noun (e.g., *Emily's prescription*). Definite NPs that begin with *the* are also known as definite descriptions [123]. These are very common in English; *the* is the most common word in most corpora (e.g., the Brown corpus [48]). Definite NPs usually involve specific reference.
- **Generic noun phrase**: an NP whose referent is a generic concept or class of things. Generic NPs are difficult to identify based on appearance alone: they may have no modifiers (e.g., NP_4) or, as in (2.4) and (2.6), resemble indefinite or definite NPs. In most cases, singular and plural generic NPs are interchangeable as is the case for (2.4),(2.5) and (2.6),(2.7).

(2.4) ***A good doctor** is hard to find.*

(2.5) ***Good doctors** are hard to find.*

(2.6) ***The heart** has four valves.*

(2.7) ***Hearts** have four valves.*

- **Pronominal noun phrase**: an NP that includes only a pronoun. Of the various kinds of pronouns, we focus on those that are personal (e.g., NP_6), possessive (e.g., *your*) or reflexive (e.g., *herself*).

  Some pronouns, called **pleonastic** or **dummy** pronouns, only fill syntactic gaps and do not actually refer to anything. In English, this is often the case for *it*, as for all three instances in example (2.8).

  (2.8) ***It**'s true that **it** was about time for **it** to rain.*

  Alternative terms sometimes used for these pronouns include 'non-anaphoric', 'non-referential', 'structural', 'expletive' and 'semantically empty'.

Referring expressions can also be **deictic**, where the reference involves something in the context of the writer or speaker (e.g., *Bring **that** over **here**.*). This kind of **exophoric** (extra-linguistic) reference is considered out of scope for traditional coreference resolution.

2.1.2 Coreference and Anaphora

As introduced in Chapter 1, referring expressions corefer if they refer to the same entity, i.e., share the same referent. When multiple expressions corefer, it is often useful to visualize them as a **coreference chain** such as {*conjunctivitis*, *pink eye*, *it*, *the condition*} from passage (2.9).

(2.9) *If you develop **conjunctivitis** (commonly known as **pink eye**) you should see your doctor in case **it** is a form that requires medication. Fortunately, **the condition** is unlikely to cause long-term damage.*

A referring expression that 'points' to another (usually preceding) expression is called an **anaphor** and the one being pointed to is called its **antecedent**. This relationship is known as **anaphora**.

Under the strict definition of anaphora, an expression is only an anaphor if it cannot be interpreted independently from its antecedent. Returning to (2.9), this is the case for *it* and *the condition*: both rely on their antecedent, *conjunctivitis*, for their meaning. We therefore consider them to be **anaphoric**. The expression *pink eye*, on the other hand, has no such dependency, so it is not anaphoric according to this strict definition. Nevertheless, *pink eye* still corefers with *conjunctivitis* (as well as *it* and *the condition*) because they refer to the same thing.

We now introduce several important classifications of anaphora. The first is in terms of the position of the antecedent relative to the anaphor. The rare situation where the antecedent comes later in the text (i.e., the anaphor 'points forward'), as in (2.10), is called **cataphora**. Generally speaking, only pronominal anaphors allow cataphora.

(2.10) ***She*** *didn't yet realize it, but* ***Heather*** *was pregnant with twins.*

Using the common definition of cataphora, as we do in this work, only discourse-new referents count. The weak definition of cataphora, as used in [63], includes all forward reference to antecedents in the same sentence, regardless of whether or not the referent is new.

The blanket term for anaphora and cataphora, 'endophora', is rarely used in practice. Most researchers prefer the simplicity of 'anaphora' where distinguishing between the two is unnecessary, and we also follow this convention.

Similarly, it is sometimes useful to distinguish between anaphors that are **intra-sentential**, where a single sentence contains both anaphor and antecedent, and those that are **inter-sentential**, where the two are separated by a sentence boundary. The anaphor in (2.10) is intra-sentential, for example. Anaphors that are pronouns tend to be intra-sentential whereas those that are nouns are more often inter-sentential; this is because pronouns are not as explicit and therefore less effective at referring to distant antecedents.

So far, we have discussed only **identity-of-reference** anaphora, where the referents involved are in fact identical. **Identity-of-sense** anaphora, on the other hand, only involves the same *kind* of referent. For example, the referring expressions in (2.11) do not actually refer to the same shirt.

(2.11) *Would you believe that shortly after Maureen bought* ***a bamboo shirt****,
Jonathan bought* ***one*** *too?*

In this thesis, we only tackle identity-of-reference anaphora, where the bulk of existing research lies. Strictly speaking, the identity-of-sense variety is not true coreference.

We can also classify anaphora by the type of relation that links anaphor and antecedent. Previous examples have only involved the identity relation, but we now introduce several others:

- Direct

  (2.12) Identity: *trans fat ... it*

  (2.13) Synonymy: *a doctor ... the physician*

  (2.14) Generalization: *leukemia ... the disease*

  (2.15) Specialization: *three cities ... Cambridge, Kitchener and Waterloo*

- Indirect

  (2.16) Set/subset: *bike helmets ... bike helmets that are cracked*

  (2.17) Class/instance: *antibiotics ... tetracycline and erythromycin*

  (2.18) Part/whole: *the appendix ... the body*

As resolving indirect anaphora often requires world knowledge [92], resolution algorithms typically focus on direct anaphora, especially the identity relation. For the purposes of this thesis, we restrict our attention to direct anaphora, which is by far the best understood.

Coreference and anaphora are related but not equivalent relationships. For instance, coreference is symmetric but anaphora is not; as we saw, an anaphor depends on its antecedent, but not vice versa [139].

These relationships are also resolved differently. Anaphora resolution is the task of identifying the antecedent of every anaphoric expression (e.g., *it* and *the condition* in (2.9)). Pronominal anaphora resolution (also known simply as 'pronoun resolution') is a restricted form of this that only handles anaphors that are pronouns. Coreference resolution is the more difficult task of identifying all coreference chains (i.e., sets of expressions that corefer) in a text. So for (2.9), coreference resolution involves finding the following chains:

1. {*you*, *you*, *your*}
2. {*conjunctivitis*, *pink eye*, *it*, *the condition*}
3. {*your doctor*}
4. {*a form that requires medication*}
5. {*medication*}
6. {*long-term damage*}

In this thesis, we focus on coreference resolution. For convenience, we sometimes apply the terms 'anaphor' and 'antecedent' to coreferential referring expressions that may not be anaphoric according to the strict definition given earlier[2]. An example would be calling *pink eye* an anaphor and *conjunctivitis* its antecedent despite the fact that *pink eye* is independently interpretable. This simplification allows us to consistently refer to all elements[3] in a coreference chain as anaphoric regardless of whether they happen to be a pronoun, definite NP, proper name, etc.

[2] Many papers use 'coreference' and 'anaphora' interchangeably [73], so this is nothing new.

[3] All elements except the first, which precedes all others and therefore cannot have an antecedent.

2.2 Coreference Resolution Strategies

As we saw for example (1.3), sometimes even humans have difficulty with anaphoric ambiguity. However, such situations are rare thanks to a number of resolution strategies that we unconsciously employ. Unfortunately, cognitive research has yet to reveal how exactly they work:

> "Whether or not these strategies are used to guide an explicit search process, or to exclude items for a search set, or both, or even to avoid an explicit search altogether, is at present unclear." [61]

These strategies, which can be divided into constraints on coreference and heuristics for its resolution [18], have proven to be essential for many approaches to automatic coreference resolution over the years.

2.2.1 Constraints

The most easily identifiable resolution strategies are syntactic constraints. Consider the following example pairs demonstrating the role that agreement in gender, number, case and person play in resolving anaphora:

- Gender

  (2.19) ***Alisa****'s dad knows* ***her*** *driving can be scary.*

  (2.20) ***Alisa's dad*** *knows* ***his*** *driving can be scary.*

- Number

  (2.21) ***Mike*** *suspected that his guild members were up to something but* ***he*** *kept quiet.*

  (2.22) *Mike suspected that* ***his guild members*** *were up to something but* ***they*** *kept quiet.*

- Case

  (2.23) *When Chandler is extra cranky,* ***Joe*** *sometimes imagines making* ***himself*** *disappear.*

  (2.24) *When* ***Chandler*** *is extra cranky, Joe sometimes imagines making* ***him*** *disappear.*

- Person

  (2.25) *Returning from a long stay with their grandparents,* ***Aaron and Hannah*** *told their parents, "Don't worry,* ***we*** *love you, too."*

  (2.26) *Returning from a long stay with their* ***grandparents****, Aaron and Hannah told their parents, "Don't worry,* ***they*** *love you, too."*

Because these agreement constraints are reliable and relatively easy to check, most resolution algorithms take advantage of them. However, such constraints are not infallible (cf. [65]). In English, for example, so-called 'singular *they*' is often used to avoid a gender-specific pronoun (or circumlocution such as *he or she*) when referring to someone of unknown gender, as in (2.27). Another case of number disagreement is (2.28), where a plural pronoun is used to refer to a singular collective noun.

(2.27) *If* ***a family member*** *has to miss the meal, make up a plate of food that* ***they*** *can reheat in the microwave later.*[4]

(2.28) *Today's fast-paced lifestyle may have you wondering how well* ***your family*** *is eating and whether you can help* ***them*** *eat better.*[4]

Semantic constraints, on the other hand, are generally not so easy to exploit computationally. One example of such a constraint is animacy, the degree to which an entity is alive. In (2.29), the inanimate pronoun *It* must refer to *the book* because the only other alternative is *Thelma*, a proper noun representing an animate entity. If the pronoun were animate, as in (2.30), then its referent would be Thelma[5].

(2.29) *Thelma told us about* ***the book****.* ***It*** *explained what life was like for American pioneers.*

(2.30) ***Thelma*** *told us about the book.* ***She*** *explained what life was like for American pioneers.*

Other semantic constraints include verbs having certain selectional restrictions, such as *drink* requiring a drinkable object. Thus the *it* in (2.31), for example, cannot refer to *the beach* because beaches are not drinkable. Just like syntactic constraints, however, such restrictions can also be violated. For instance, in (2.32), metaphor allows the referent to be Cylena's Jeep despite the fact that vehicles do not normally drink.

[4] From "Managing Family Meals" [104].

[5] Note that gender agreement could also be used in this case. However, relying exclusively on gender can be problematic, especially for ambiguous (e.g., Pat, Sam, Drew) or unknown names.

(2.31) *Matty gets another* ***mai tai****, stretches out on the beach, and drinks* ***it*** *while enjoying the sunset.*

(2.32) *Cylena loves her* ***Jeep****, but* ***it*** *drinks gasoline like you wouldn't believe.*[6]

Even with complete understanding of animacy, selectional restrictions and metaphor, a coreference resolution algorithm would still frequently fail. In fact, it is often not linguistic knowledge at all but world knowledge that is required to resolve coreference, as in the following classic set of examples:

(2.33) *The* ***monkey*** *ate the banana because* ***it*** *was hungry.*

(2.34) *The monkey ate the* ***banana*** *because* ***it*** *was ripe.*

(2.35) *The monkey ate the banana because* ***it*** *was lunch time.*

For the first sentence, the resolver needs to know that monkeys can be hungry, while bananas cannot; for the second, that monkeys are not generally considered ripe but bananas are; and for the third, that lunch time is when eating often happens.

Ultimately, almost any piece of shared knowledge could turn out to be relevant in resolving coreference [69].

2.2.2 Heuristics

There are also a number of heuristics (sometimes called 'preferences' [18] or 'indicators' [88]) that can be leveraged in resolving coreference. While not as dependable as the aforementioned constraints, these heuristics are nevertheless important factors that are frequently exploited by existing algorithms.

2.2.2.1 Recency

In discourse, recently mentioned entities are more salient than earlier ones. Proximity is also a factor: as the gap between anaphor and antecedent widens (i.e., the number of intervening words increases), anaphora resolution becomes more difficult [102]. In passage (2.36), for example, *it* could refer to *a flower*, *a dress* or *a platypus*, but the latter is most probable because of the effect of recency.

(2.36) *Brad bought Candice a flower. Later, he bought her a dress. Finally, he bought her a platypus.* ***It*** *was definitely her favourite gift.*

[6] Example adapted from [69].

Focus is also relevant in this context (cf. attentional state [58]). Entities that have been explicitly mentioned in recent discourse are said to be 'in focus' or 'foregrounded' and can be referred to pronominally. (Otherwise, a full noun phrase might be necessary.) Entities will remain in focus (and working memory) if they are part of the theme of the discourse. Studies show that pronouns are processed more quickly when the referent in question is still in focus [53].

2.2.2.2 Repetition

Frequently mentioned or reiterated entities are more likely to be referred to anaphorically [88], sometimes even taking precedence as potential antecedents over recently introduced entities. In passage (2.37), notice that newcomer Danielle does not become the preferred antecedent. The effect of repetition is related to working memory and focus.

(2.37) ***Jessica*** *went into the pharmacy but then realized that* ***she*** *forgot* ***her*** *purse.* ***She*** *went back and grabbed it from the van. Danielle teased* ***her*** *for forgetting.* ***She*** *wisely decided not to retaliate.*

2.2.2.3 Syntactic role

There is also a bias towards referring to entities found in prominent positions of a sentence such as subject and direct object [27, 50]. In (2.38) and (2.39), notice that switching the syntactic roles of Andrew and Mark causes a corresponding change in interpretation.

(2.38) ***Andrew*** *was shooting hoops with Mark.* ***He*** *sunk three consecutive three-pointers.*

(2.39) ***Mark*** *was shooting hoops with Andrew.* ***He*** *sunk three consecutive three-pointers.*

Nouns within prepositional phrases tend to fill minor syntactic roles (e.g., indirect object). For this reason, non-prepositional phrases are more likely to contain antecedents [88].

2.2.2.4 First mention

Evidence from eye tracking experiments suggests that cognitive structure building processes make the first-mentioned noun phrase the preferred antecedent of an ambiguous pronoun [57]. Note that this heuristic may conflict with the previous one, as in the following example:

(2.40) *After dropping Leila off,* ***Lacey*** *headed straight home.* ***She*** *was exhausted.*

In this case, the first mention heuristic fails: *Leila* is not the preferred antecedent of *she* despite being mentioned first. This seems to suggest that syntactic role takes priority over first mention. Actually, recent psychological experiments indicate that both first mention and subject preference have an effect, i.e., that pronoun resolution is determined by a delicate interplay of several factors [68].

2.2.2.5 Verb causality

Some verbs seem to have an implicit causality whereby one argument receives more emphasis than the other(s). For example, *defeat* is an 'NP_1 verb' whereas *accuse* is an 'NP_2 verb', as illustrated by (2.41) and (2.42). The latter case demonstrates that causality can overpower subject preference.

(2.41) ***Noah*** *defeated Murray because* ***he*** *cheated.*

(2.42) *Noah accused* ***Murray*** *because* ***he*** *cheated.*

Verb causality is also known as 'causal valence' [65].

2.2.2.6 Parallelism

In general, there is a preference for matching anaphors to antecedents in parallel positions [126]. Consider the following example:

(2.43) *Brad took* ***Jordan*** *to an Edmonton Oilers game. James took* ***him*** *to a Calgary Flames game.*

As the subject of the first sentence, *Brad* would normally be the most likely candidate antecedent of *him*. However, parallelism takes precedence over syntactic role in this case.

Hirst [65] has a different perspective on parallelism, which he illustrates as follows:

(2.44) ***Ross*** *likes* ***his*** *beer and* ***Daryel his*** *carrot juice, but* ***Bruce*** *swears by* ***his*** *Samoa Fogcutter (two parts gin, one part red wine).*

(2.45) ***Roger*** *makes some great drinks at home. Ross likes* ***his*** *beer and Daryel* ***his*** *carrot juice, but Bruce swears by* ***his*** *Samoa Fogcutter.*

In the first example, each *his* refers to the preceding name whereas the extra context in (2.45) causes them all to refer back to Roger. The parallelism is that, in both cases, all the pronouns behave in a consistent way.

2.3 Related Work

This thesis may tread unfamiliar ground, but there has been considerable research in related areas. We first survey existing corpora that are related to our own and then introduce the problem of reliably annotating coreference. Finally, we provide an overview of previous approaches to coreference resolution.

2.3.1 Existing Corpora

Despite their importance to the field, there are still only a handful of corpora annotated for full coreference[7] (Table 2.1). This is perhaps surprising considering the vast amount of work in the area and the number of tools that support coreference annotation (see Section 2.3.2).

In some cases, the annotated corpora are not even publicly available (e.g., the Lancaster Anaphoric Treebank [54], which was commercially funded). Over 80% of the annotated data is news-related (e.g., *Wall Street Journal*, *New York Times*) and the rest is technical manuals and narrative.

Several health corpora are known to exist, including recent parallel corpora. The Pan American Health Organization corpus, for instance, is a Spanish-English parallel corpus of 180 documents on Latin American health issues [29]. Another is the Chinese-English Parallel Health Corpus, which consists of 31,638 words from public health pamphlets by the British Government [84]. Recent work by Deléger et al. [32, 33] involved building a parallel French-English corpus of over 27 million words downloaded from the Health Canada website. This latter work motivated us

[7] We do not count the datasets from the Automatic Content Extraction (ACE) evaluations (e.g., [38]), which are restricted to coreference among certain entities such as PERSON and ORGANIZATION.

Table 2.1: Known coreferentially annotated corpora

Corpus	Words
Lancaster Anaphoric Treebank [54]	100,000
Penn Treebank [56]	94,500
Message Understanding Conferences (MUC)[8] [22, 23]	60,000
Downloaded technical manuals [96]	55,444
Jules Verne's *From the Earth to the Moon* [96]	4,965
Total	314,909

to investigate resources available from Health Canada for inclusion in our health education corpus (see Section 3.1).

Another dimension of corpus-based research has focused on health care discourse, especially doctor-patient interactions (e.g., [128, 129, 137]). More recently, researchers have been investigating other varieties of health discourse. The NHS Direct corpus, for example, includes 61,981 words of transcribed phone calls made to a health advice and information service in England [1].

So while health corpora do exist, none focus on consumer-oriented health education. Furthermore, none have been manually annotated for coreference.

2.3.2 Coreference Annotation

As shown in the previous section, few corpora have been annotated for full coreference. The reasons for this no doubt include the difficulty of reliably annotating coreference and the labour-intensive nature of the task (as discussed, for example, in [95]). The task has (to some degree) been simplified by the development of numerous annotation tools over the years. These tools include the following:

- DTTool [2]
- Alembic [31]
- MATE [85]
- MMAX [101]
- CorefDraw [60]
- GATE [28]
- PALinkA [105]
- WordFreak [98]

[8] The coreference task was not incorporated until MUC-6 and MUC-7. The combined amount of data annotated for these conferences has been reported as both 54,800 [130] and 65,000 [90], so we take the average.

Unfortunately, the task remains challenging even with the support of such tools.

In actuality, annotating coreference is not difficult. *Reliably* annotating coreference is the problem. One way to determine the quality of annotation is for multiple people to perform overlapping annotations, which can then be compared. This allows computing inter-annotator agreement, i.e., the degree of consensus among annotators.

Early work involving discourse annotation used various measures like percent agreement for determining inter-annotator agreement. In the mid-90s, Carletta [19] proposed that the kappa (κ) statistic [127] (cf. [25]) used in content analysis be adopted as a standard measure of reliability to ensure comparability of results in discourse. The main advantage of κ is that it factors in the likelihood of annotators agreeing by chance. As κ became a de facto standard in the discourse community, several issues were raised such as its appropriateness when more than two annotators are involved or for certain agreement tasks (e.g., [34, 81]).

For the specific task of coreference annotation, a scoring scheme based on precision and recall was developed as part of MUC-7 [143]. While the MUC scoring scheme is still used today (e.g., for the recent Anaphora Resolution Exercise [106]), the κ statistic has been favoured because it takes expected agreement into account [111]. Unfortunately, κ uses a binary notion of agreement that treats partial agreement (e.g., {*Dr. Smith, he, the doctor, he*} and {*Dr. Smith, he, he*}) as total disagreement. Because it is common for annotators to overlook at least some referring expressions (in fact, this seems to be the most common issue [64]), κ tends to be very low for anaphora and coreference annotation.

The alpha (α) statistic addresses the partial agreement problem by incorporating a distance metric [78]. The most common metric used for coreference is that introduced by Passonneau [112], though Jaccard and Dice's information retrieval metrics [80] are sometimes provided for comparison (e.g., in [114]).

For both κ and α, values range from 0 (equivalent to chance agreement) to 1 (perfect agreement). The threshold for moderate agreement is usually considered to be 0.67 [78], although this cut-off has been criticized and other scales exist [39].

2.3.3 Coreference Resolution

Not unlike most areas of natural language processing, coreference resolution is far from being solved. As already illustrated, this task is especially dependent on world knowledge and, in some cases, even such knowledge is not enough.

It did not take long for researchers (e.g., [20]) to realize the complexity of the task, and early approaches did not hesitate to exploit available knowledge sources such as grammars, lexicons and discourse models. More recently, the trend has been toward stripped-down approaches that eschew complex knowledge sources and processing in order to focus on speed and reliability. We refer to the former as **knowledge-rich** approaches and the latter as **knowledge-lean**.

Considering the vast amount of work in coreference resolution over the past three decades, it would be counter-productive to attempt an exhaustive review of the literature. Instead, we emphasize algorithms that, like our own, adopt the knowledge-lean approach. For an in-depth overview of all research to date, *Anaphora Resolution* [91] is highly recommended.

The naïve pronoun resolution algorithm of Hobbs [66] traverses parse trees in a specific order while searching for antecedents. The approach requires full syntactic parsing and checks for gender, number and person agreement. Hobbs' naïve algorithm set the standard for later syntactic approaches and has frequently been used as a baseline during evaluations (e.g., [6, 56, 79, 96]). It also began the unrealistic trend of evaluating algorithms on only third-person pronouns with perfect preprocessing and no instances of pleonastic *it*.

The Resolution of Anaphora Procedure (RAP) developed by Lappin and Leass [79] continued in the use of syntactic parsing but also incorporated weighted 'salience factors', including the recency and syntactic role heuristics (covered in Section 2.2.2). Another of RAP's contributions to later work was introducing pattern-based pleonastic pronouns filtering.

One of the first truly knowledge-lean resolution systems was that of Kennedy and Boguraev [71], an adaptation of RAP that replaced the full syntactic parse with part-of-speech tagging and grammatical function. Their evaluation revealed only a small decrease in performance despite limited knowledge.

Baldwin [6] developed CogNIAC, which continued the trend of only relying on morphological and shallow syntactic information initiated by Kennedy and Boguraev. Unlike previous approaches, CogNIAC used a series of ordered rules designed for high precision. These rules allowed the algorithm to avoid making guesses in highly ambiguous cases; the downside is that some pronouns were not resolved at all.

Another landmark knowledge-lean pronoun resolver was Mitkov's robust approach [88]. Like CogNIAC, this system required only the output of a part-of-speech tagger and NP extractor. Antecedents were selected based on a score computed from a variety of 'antecedent indicators', which cover heuristics such as recency, repetition and syntactic role. Mitkov's system is similar to RAP [79] in that the indicators resemble salience factors, and both systems were evaluated on technical manuals. A later version of this approach, dubbed MARS [94], included pleonastic *it* and animacy detection, and also operated in fully automatic mode (i.e., without hand-edited preprocessing).

The final pronoun resolution system we considered was the light-weight approach of Dimitrov et al. [37]. This approach was similar to the others except being restricted to cases where the antecedent is a named entity (e.g., person, organization, location). The light-weight algorithm is notable in that it attempted to resolve more than just third-person pronouns and included improved pleonastic *it* detection (relative to that of RAP [79]).

Systems attempting full coreference (not just pronoun) resolution became more popular after the coreference task was added to MUC-6 and MUC-7 [99, 100]. The best score on the MUC-6 Coreference Task was achieved by the resolution algorithm [70] included in FASTUS [4]. This coreference resolver was fully automatic and knowledge-lean and handled proper names and definites in addition to pronouns. The approach also included name alias recognition (e.g., *GM* for *General Motors*). The top scorer from MUC-7 was LaSIE-II [67], which adopted a knowledge-rich approach that included chart parsing and discourse interpretation. Like previous approaches, the algorithm used a similarity score and various constraints.

Supervised learning was applied to pronoun and coreference resolution fairly early (e.g., [3, 26]), but it was Soon et al. [130] that first managed to achieve scores that were competitive with heuristics-based approaches on identical data. We do not further consider learning-based approaches in this thesis for the following reasons:

1. Our resolution algorithm is not learning-based. (Our position is that it would be inappropriate to apply machine learning to a new domain before even investigating basic heuristic approaches.)

2. Exploration, not optimization, is our primary objective. Learning-based approaches are less transparent than heuristics-based ones, and their output is typically more difficult to interpret.

3. We are working in a new domain for which limited data is available.

A substantial number of previous approaches target specific varieties of anaphora. This is particularly evident with pronouns, but also true for definite descriptions (e.g., [55, 140], named entities (e.g., [37, 83]) and even pleonastic *it* (e.g., [40, 41, 108]).

In this thesis, we tackle all of these kinds of anaphora[9]. In other words, our algorithm is designed for full coreference resolution. We also handle all personal, possessive and reflexive pronouns (regardless of grammatical person), and do not manually correct preprocessing or remove instances of pleonastic *it* as was done for many of the above approaches.

[9] That is, all identity-of-reference direct nominal anaphora, as described in Section 2.1.

Chapter 3

Creating a Health Education Corpus

In order to determine whether coreference phenomena are significantly different in the health education domain compared to previously studied domains, we first require a representative corpus. Unfortunately, no suitable health education corpora are known to exist, and relevant samples in existing corpora are rare.

Popular large-scale corpora incorporate language from a variety of sources in order to better represent the overall use of a given language. These general corpora, however, typically contain only a small number of medical samples, and even then these samples are largely practitioner-oriented—too technical for patients and related consumers of health information. For example, the Brown corpus [48] contains a mere five texts in the 'medicine' category. The British National Corpus (BNC) [15] has significantly more, but the majority are excerpts from scientific journals (e.g., *British Medical Journal* and *Nucleic Acids Research*).

To address this need, we created a new corpus of health education texts. The creation process involved four main steps: sample selection, encoding, automatic part-of-speech annotation and manual coreference annotation. Each step is discussed in turn in the following sections.

Given the lack of such specialized corpora, we plan to make this health education corpus, including the manually annotated portions, freely available to the wider research community for future use.

3.1 Sample Selection

Unlike general corpora, this corpus does not attempt representation of a language as a whole. Instead, it is designed to capture only a specific domain, namely (English) health education. Even with such a specialized focus, exhaustive representation is impossible. We defined the following desiderata to guide the selection of samples for inclusion in the corpus:

D1. **Free**: The information should be freely available to the public. Freedom of reproduction is also highly desirable.

D2. **Reliable**: Although not intended to be used by actual patients, the information should be of sufficient quality to be suitable for such use.

   (a) **Objective**: To avoid bias, information from credible sources without conflicts of interest (e.g., commercial interests) such as government and non-profit organizations are given precedence.
   (b) **Accurate**: Information provided should be factual as per the present state of medical knowledge.
   (c) **Current**: Related to above, information must be up-to-date. We consider material written or updated after 2000 as current.

D3. **Consumer-oriented**: The material must be written in such a way as to be easily understood by non-technical readers, but not over-simplified to the point of inaccuracy.

The Web is widely recognized as one of the largest sources of consumer health information. Surveys indicate that approximately 80% of connected Americans have sought health information online at some point [47]. In fact, health information is some of the most widely sought content on the Internet, comprising about 4.5% of all searches [42]. Because the World Wide Web is an open resource, health information found there satisfies D1, but may conflict with D2: the information could be from untrustworthy or ill-informed sources instead of medical professionals. To avoid this issue, we carefully selected the sources of included text.

Following Deléger et al. [33], we selected the Health Canada website as one source, although we are exclusively interested in the It's Your Health (IYH) series of articles [17]. These informative articles are written specifically for consumers by scientists and experts from Health Canada and the Public Health Agency of Canada, and cover a broad range of health-related topics:

- Diseases & Conditions
- Environment
- Food & Nutrition
- Lifestyle
- Medical Information (e.g., treatments)
- Consumer Products

EatRight Ontario (ERO) [104] was selected as the second source of health information. This website contributes persuasive articles exclusively about nutrition and healthy eating, balancing the primarily informative and broad nature of the IYH material. The ERO information is both current and reliable, provided by the Government of Ontario's Ministry of Health Promotion in partnership with Dietitians of Canada.

Finally, online decision guides from the Mayo Clinic [24] were selected for inclusion, providing a mix of information and personal testimonial about specific treatments. These popular guides are easy-to-read and accurate, written by medical experts at the Mayo Clinic, a not-for-profit medical practice. They cover the advantages and disadvantages of specific decisions such as birth control methods and mastectomy versus lumpectomy as breast cancer treatment. It should be noted that these guides are much longer than the average health education article, but there are also far fewer of them. Furthermore, linguistic style varies considerably between sections, but this is not uncommon in medicine [14].

A sample text from each of the three sources is included as Appendix A.

3.2 Encoding

After being selected, the online material was downloaded[1] from each source. Some webpages that were not full articles (e.g., printable charts and recipes) were discarded. For the remaining articles, everything but the actual article content was stripped away (headers, navigation menus, tables of contents, footers, glossaries, etc.), and some minor encoding issues were manually corrected. The decision was made not to strip HTML tags because they provide useful structural information (e.g., the location of headings).

The basic structure of the collected samples was then manually marked-up in XML according to the Text Encoding Initiative's Guidelines for Electronic Text Encoding and Interchange [131]. This standardized format facilitates re-use and, being XML, allows automatic validation of document structure. Some of the markup tags from the TEI guidelines are as follows:

`<body>...</body>` - body of a text

`<p>...</p>` - paragraph

`<s>...</s>` - sentence

`<w type="NN">...</w>` - grammatical word and its part-of-speech

`<ref type="coref" target="12"/>` - reference to an antecedent

Note that some of these tags (i.e., the final three) are not used until later linguistic annotation is performed (as described in subsequent sections).

[1] Where convenient, an open-source web crawler called WebSPHINX [87] was used.

Another advantage of this encoding is the availability of numerous XML-based tools such as Extensible Stylesheet Language Transformations (XSLT) and the XML Path Language (XPath), both of which proved useful while performing analysis and resolution on the corpus. Another benefit of the XML encoding is being able to use Xaira (XML Aware Indexing and Retrieval Architecture), an advanced version of SARA, the concordancing software developed for searching the BNC [15].

3.3 Part-of-Speech Annotation

Once encoded in XML, the corpus was annotated for part-of-speech (POS) information using a pipeline of modules included with GATE (General Architecture for Text Engineering) [28]:

1. Tokenization
2. Sentence splitting
3. Part-of-speech tagging

We opted to use TreeTagger [124], which is based on decision trees, for POS tagging because it was experimentally found to perform better on this corpus than the default tagger included with GATE. Previous research has shown that TreeTagger handles unknown (e.g., domain-specific) terms gracefully even in the medical context [113]. It uses the Penn TreeBank [82] tagset with a slight refinement for verbs. Sample output[2] is shown in Figure 3.1.

In total, the health education corpus consists of 339,027 words[3] in 230 articles (Table 3.1).

3.4 Coreference Annotation

Coreference annotation is a notoriously labour-intensive task [97]. As current automatic techniques do not adequately handle coreference phenomena, manual analysis was necessary. Using a newly developed annotation scheme (described in Section 3.4.1), a subset of the health education corpus was manually annotated for full coreference chains as described in Section 3.4.2. This work was performed by two annotators to allow reliability statistics to be computed (see Section 3.4.3), helping ensure reproducibility of results.

[2] From “Asthma” [17].

[3] Excluding punctuation, though the tagger treats punctuation as words.

Asthma is a chronic lung disease that can be fatal.

```
<s>
  <w type="NP" kind="word" lemma="Asthma" length="6">Asthma</w>
  <w type="VBZ" kind="word" lemma="be" length="2">is</w>
  <w type="DT" kind="word" lemma="a" length="1">a</w>
  <w type="JJ" kind="word" lemma="chronic" length="7">chronic</w>
  <w type="NN" kind="word" lemma="lung" length="4">lung</w>
  <w type="NN" kind="word" lemma="disease" length="7">disease</w>
  <w type="WDT" kind="word" lemma="that" length="4">that</w>
  <w type="MD" kind="word" lemma="can" length="3">can</w>
  <w type="VB" kind="word" lemma="be" length="2">be</w>
  <w type="JJ" kind="word" lemma="fatal" length="5">fatal</w>
  <w type="SENT" kind="punctuation" lemma="." length="1">.</w>
</s>
```

Figure 3.1: A sample sentence and its XML representation in the corpus

Table 3.1: Breakdown of the health education corpus

			Words/text		
Source	Texts	Words	Mean	Min	Max
EatRight Ontario	67	53701	801.51	267	3047
It's Your Health	148	132600	895.95	436	1714
Mayo Clinic	15	152726	10181.73	4362	18786
Total	230	339027	1474.03	267	18786

Because of the difficulty of the task, manually annotating coreference in all 230 articles would not be feasible. To determine how many texts to annotate, we concentrated on the number necessary to adequately represent each source, which Biber [13] found to be approximately ten (using 1,000 word samples). In some cases, our sample sizes had to be considerably smaller because the articles themselves are short, especially those from EatRight Ontario.

To prevent any single source from having too great an influence, samples were selected such that each source contributes approximately one third of the subcorpus[4]. The result is a balanced representation of the health education genre (including general information, persuasive articles and testimonials). Figure 3.2 lists all samples included in the manually annotated subcorpus.

In total, the manually annotated subcorpus of health education articles contains 28,505 words (4,120 instances of coreference) across 39 texts[5] (Table 3.2). The annotation of full coreference chains took approximately 60 person-hours[6], an average of one hour and 22 minutes per document. The approximate annotation speed was 525 words/hour.

Table 3.2: Breakdown of the manually annotated subcorpus

			Words/text		
Source	Texts	Words	Mean	Min	Max
EatRight Ontario	17	9711	571.24	305	806
It's Your Health	13	10532	810.15	557	1145
Mayo Clinic	9	8262	918.00	484	1160
Total	39	28505	730.90	305	1160

[4] The Mayo Clinic material includes between two to five testimonials for most of the decision guides. To avoid over-representation among these guides, only one testimonial was selected from each. A negative side-effect is that the overall word count is slightly lower.

[5] The total is actually 31,529 words (4,598 instances of coreference) across 44 texts, taking into account that some documents were annotated twice, once by each annotator, to allow inter-annotator reliability to be computed.

[6] The reported time does not include training time (i.e., learning the software and the annotation scheme) or even short breaks—the tool used includes its own timer, which was paused when breaks or discussion were necessary.

Words	Title
	EatRight Ontario
689	Are You A Yo-Yo?
305	Breastfeeding is best for baby
641	Children's Health - Overweight and Obesity
708	Decoding the New Nutrition Label
616	Eating right at school
533	Facts on Fluids - How to stay hydrated
617	Food Safety - True or False
403	Make a balanced breakfast a habit in your home
760	Managing Family Meals
609	Meal makeovers that pack more veggies and fruit
806	Pass The Bread
405	Portion Distortion
728	Pre-pregnancy healthy eating checklist
377	Probiotics - The Good Bacteria
514	Simple lunch solutions
604	Tackling Trans Fat
396	Tips for Healthy Digestion
	It's Your Health
1145	Acne Treatments
847	Assessing and Managing the Health Risks of Living Biotechnology Products
774	Asthma
986	Candle Safety
803	Chickenpox Vaccine
632	Chlamydia
557	Garlic-In-Oil
980	Halloween Safety
1000	Medical Test Kits for Home Use
751	Responsible Holiday Drinking
671	Safety of Exposure to Electric and Magnetic Fields from Computer Monitors
666	Screening for Colorectal Cancer
720	Selling Second-hand Products
	Personal Stories from Mayo Clinic Decision Guides
847	Adjuvant therapy for breast cancer guide
1160	Back pain guide
820	Carpal tunnel syndrome guide
861	Depression treatment guide
484	Ear infection guide
1106	Enlarged prostate (BPH) guide
1112	Prostate cancer guide
1067	Uterine fibroids guide
805	Vaginal birth after C-section (VBAC) guide

Figure 3.2: Complete list of texts in the manually annotated subcorpus

3.4.1 Annotation Scheme

We use a custom annotation scheme derived from that described in the Coreference Task Definition of MUC-7 [100]. The MUC-7 annotation scheme was chosen over the many alternatives (e.g., UCREL [54], MATE [115]) for the following reasons:

1. The scheme's popularity, facilitating comparison of results.
2. Its emphasis on quality and reliability of annotations (as opposed to completeness of coverage).
3. Availability of compatible tools and resources (e.g., PALinkA [105]).

The MUC-7 scheme targets the most common, and best understood, kinds of anaphora: nominal, where the antecedent is a noun phrase, and identity-of-reference, where the anaphor and antecedent refer to the same (not merely similar) real-world referent.

However, the MUC-7 scheme is not without criticism (e.g., [139]). One of its limitations is covering only the relation of identity between anaphor and antecedent. Following the NP4E annotation scheme [62], we also annotate other relations involved in direct anaphora: synonymy (e.g., *doctor . . . physician*), generalization (e.g., *leukemia . . . the disease*) and specialization (e.g., *medical test kit for home use . . . the product*). Unfortunately, distinguishing between these relations is sometimes difficult. Also like NP4E, our annotation scheme explicitly identifies the kind of reference from six possibilities:

- **Appositive**: an appositive phrase (usually set off by commas)

  (3.1) *Acne, **a common skin condition**, can be treated in many ways.*

- **Copular**: a linking verb used to associate subject and predicate

  (3.2) *Nova Scotia **is** the province with the highest cancer rates in Canada.*

- **Bracketed**: an NP in brackets immediately following another NP

  (3.3) *Trigeminal neuralgia **(TN)** is a disorder of the nervous system.*

- **Dashed**: an NP separated from another NP by an em dash

  (3.4) *Many Canadians do not realize is that cancer—**the leading cause of premature death**—is mostly preventable through healthy living.*

- **Cataphoric**: reference to an expression whose first mention comes later

  (3.5) *When Dr. Ardnek insinuated **he** was illiterate, **Greg** was shocked.*

- **NP**: an NP not satisfying any of the above

  (3.6) ***Troy** may have a gluten intolerance but **he** doesn't miss a beat.*

The following sections detail other significant ways in which this annotation scheme differs from those of MUC-7 and NP4E.

3.4.1.1 Markables

A **markable** is an NP that may participate in coreference according to the annotation scheme. Most NPs are considered markable, but there are a few exceptions such as substrings of named entities (e.g., *Canada* in *Health Canada*) and prenominal modifiers (e.g., *chickenpox* in *chickenpox vaccine*)[7].

Unlike the schemes of both MUC-7 and NP4E, we do not aim to annotate all markables; instead, we require tagging only as many as necessary to properly annotate coreference. This is made feasible by treating words identified by the part-of-speech tagger as nouns (singular, plural, proper) and pronouns (personal, possessive, etc.) as suggestions for potential markables. The advantage is saving annotation effort: the two-pass task (markables, then coreference) becomes single-pass.

Our decision is supported by the observation of van Deemter and Kibble [139] that "a strict distinction between the two steps is difficult to maintain, because, in principle, almost anything is markable". Furthermore, the MUC-7 scheme already includes cases where NPs are not markable unless they actually corefer with something (e.g., prenominal modifiers and conjuncts).

3.4.1.2 Gerunds

Both the NP4E and MUC-7 schemes differentiate between gerunds which are nominalizations of verbs and those which are not. In this scheme we follow the POS tagger: if it identifies the word as a noun, we treat it as a noun. Otherwise, it is not markable. These situations are exemplified by (3.7) and (3.8), respectively.

(3.7) ***Swimming*** *is a great way to exercise.*

(3.8) ***Losing*** *weight should be a gradual process.*

Of course, if the annotator is certain that the tagger is wrong, he or she may choose to make an exception.

3.4.1.3 First and second pronouns

In existing corpora, first-person and second-person pronouns are mainly found in dialogue and reported speech. Health education, on the other hand, involves frequent use of these pronouns outside of dialogue (see Section 4.2) as in (3.9). We annotate such cases as coreferential because *you* and *your* clearly have the same referent (the reader), even if the exact identity of that referent is unknown.

[7] Unless the modifier occurs elsewhere as the head of an NP.

(3.9) *Contact **your** doctor if **you** are thinking about using an isotretinoin product and **you** or members of **your** family have diabetes.*[8]

On rare occasions such as (3.10), *you* does not refer specifically to the reader but rather to people in general. Where the distinction was clear, such cases were not annotated as coreferential with other instances of *you*.

(3.10) *As a general rule, **you** can only get chickenpox once.*[9]

3.4.1.4 Generic coreference

Health articles are full of reference to general concepts, unlike news articles, which usually report on specific events among specific people, organizations, etc. Examples of such general concepts include *symptoms, airways, triggers, health care provider, cure, treatment, research, medication* and *oral contraceptives.* Consider the following example:

(3.11) *Acne is caused by inflammation of **the oily glands in the skin**. When the duct of **the gland** becomes blocked by layers of skin, . . .* [8]

Despite the use of definite noun phrases, the mentions of *gland* would not normally be considered coreferential because they do not refer to a specific gland. Nevertheless, these expressions are both referring to a generic gland.

It could be argued that expressions referring to such general concepts[10] do not actually corefer, but we decided to annotate them because of their prevalence in the health corpus and the difficulty of knowing where to draw the boundary between generic and specific concepts. This is an outstanding issue in other schemes (e.g., NP4E [62]), where they are sometimes marked but other times not (perhaps depending on frequency of mention). Meanwhile, other schemes do not annotate generic coreference all (e.g., PoCoS [77]). How exactly this relates to identity-of-sense (as opposed to identity-of-reference) anaphora is a matter for future investigation.

3.4.2 Annotation Procedure

Before beginning annotation, annotators were given a detailed annotation guide (included as Appendix B) to familiarize them with the software and annotation

[8] From “Acne Treatments” [17].

[9] From “Chickenpox Vaccine” [17].

[10] It should be pointed out that dates, times and quantities, all of which are considered markable in news texts, could also be regarded as general concepts.

scheme. They then performed a supervised trial run during which any questions were answered.

While annotating in XML could (in theory) be done by hand, this is impractical for any significant amount of text. We instead used a specialized editor designed for this task called PALinkA [105], which greatly simplifies the annotation process by providing a graphical user interface (Figure 3.3), shortcut keys, and automatic generation and management of unique identifiers. It also includes a (supervised) automatic tagging feature, giving annotators the option of searching the text for identical strings and annotating those as well. Another useful feature of PALinkA is tracking how long it takes for each document to be annotated.

The basic annotation procedure is as follows:

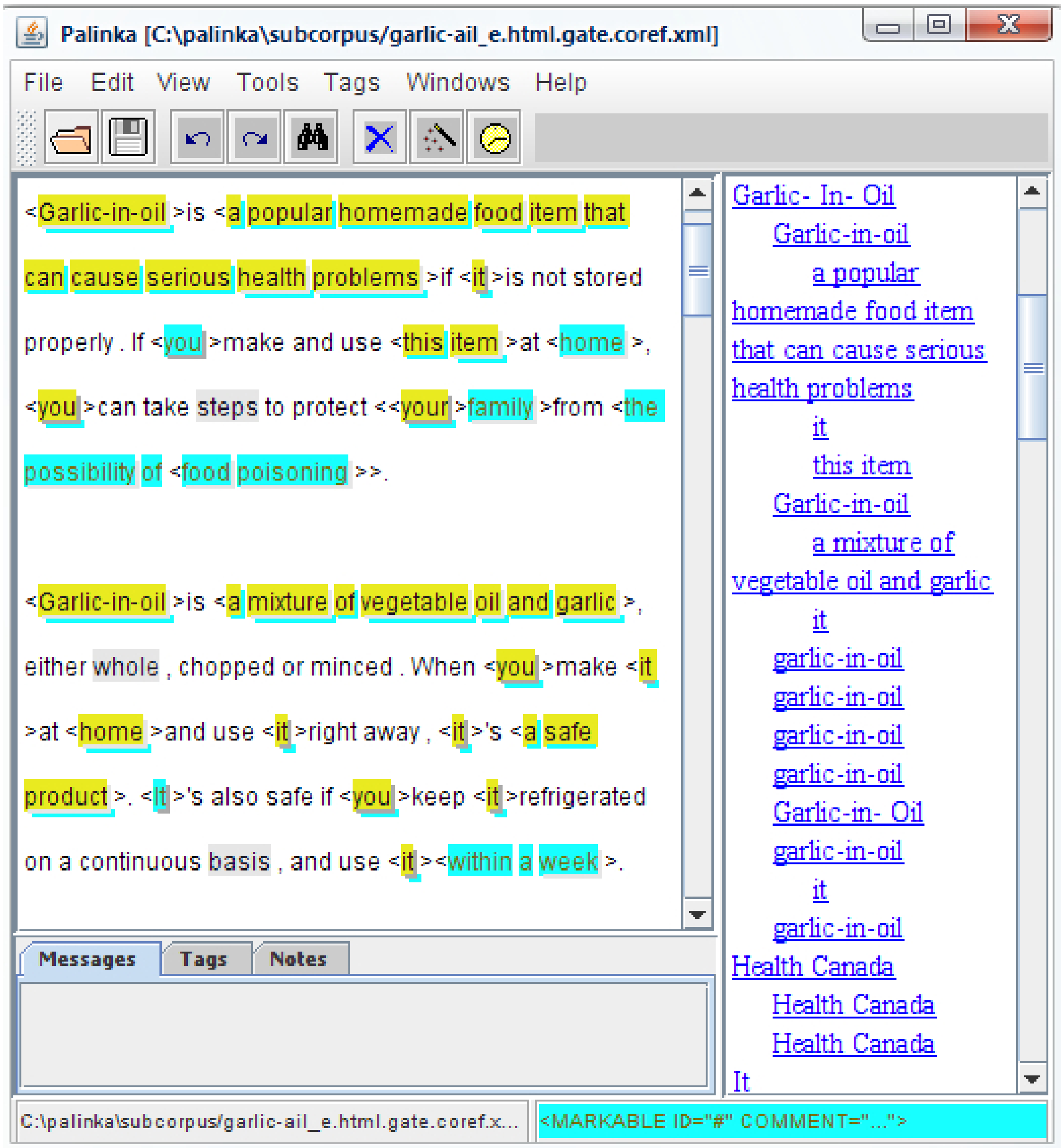
1. Open a new document in PALinkA. Noun phrases identified by the POS tagger are highlighted grey.
2. Briefly skim the document to become familiar with its content.
3. Annotate topic NPs (e.g., *asthma* in the "Asthma" article [17]) all at once using the auto-tagging feature.
4. Carefully scan through the document word-by-word looking for coreference, using the suggested NPs as potential markables. When coreference is found:
   (a) Add markable annotations to the coreferring NPs.
   (b) Add a coreference link between the new markables.
5. Once at the end of the document, check it over to ensure nothing was missed.

Annotating a markable as in step 4(a) is done as follows:

1. Select the <MARKABLE> tag if not already selected (F9).
2. Select the group of words to be annotated as a markable.
3. Press ENTER to confirm (or ESC to cancel).

Adding a coreference annotation (step 4(b)) is similarly straightforward:

1. Select the <COREF> (or <UCOREF>) tag if not already selected (F11 or F12, respectively).
2. Click on a markable that corefers with another.


Palinka [C:\palinka\subcorpus/garlic-ail_e.html.gate.coref.xml]
File
Edit
View
Tools
Tags
Windows
Help
Messages
Tags
Notes

<Garlic-in-oil >is <a popular homemade food item that can cause serious health problems >if <it >is not stored properly . If <you >make and use <this item >at <home >, <you >can take steps to protect <<your >family >from <the possibility of <food poisoning >>.

<Garlic-in-oil >is <a mixture of vegetable oil and garlic >, either whole , chopped or minced . When <you >make <it >at <home >and use <it >right away , <it >'s <a safe product >. <It >'s also safe if <you >keep <it >refrigerated on a continuous basis , and use <it ><within a week >.

Garlic- In- Oil
Garlic-in-oil
a popular homemade food item that can cause serious health problems
it
this item
Garlic-in-oil
a mixture of vegetable oil and garlic
it
garlic-in-oil
garlic-in-oil
garlic-in-oil
garlic-in-oil
Garlic-in- Oil
garlic-in-oil
it
garlic-in-oil
Health Canada
Health Canada
Health Canada
It

C:\palinka\subcorpus/garlic-ail_e.html.gate.coref.x...

<MARKABLE ID="#" COMMENT="...">

Figure 3.3: Annotating coreference using the PALinkA tool

3. Click on the markable that corefers with the first (either directly in the text, or from the list on the right side of the interface).
4. Select values for the relationship and reference attributes (or just press ENTER to accept the default values).

The annotators were encouraged to annotate at the same time, allowing them to immediately discuss any issues or unclear cases they encountered[11]. If this was not possible or the issue could not easily be resolved, the special <UCOREF> tag was used to flag it. Obvious typos (e.g., *there* for *their*) were treated as though the typo did not occur, with a comment to that effect inserted in the annotation.

3.4.3 Inter-Annotator Agreement

Overlapping annotation was performed on a 10% sample[12] across all parts of the manually annotated subcorpus in order to determine inter-annotator agreement (Table 3.3). Of the various reliability measures available for discourse annotation (see Section 2.3.2), we use precision and recall as per the MUC scoring scheme [143] along with Krippendorff's α statistic [78]. While once favoured in coreference annotation (e.g., [111]), the κ statistic lacks support for partial agreement—almost inevitable in coreference annotation [64]—and so has not been computed.

Table 3.3: Inter-annotator agreement in terms of precision, recall and the α statistic (using various distance metrics)

					Krippendorff's α			
Text	Words	P	R	F	Pass.	Jacc.	Dice	Avg.
Carpal Tunnel (MC)	820	0.898	0.888	0.893	0.787	0.732	0.803	0.774
Colorectal (IYH)	666	0.843	0.816	0.829	0.673	0.630	0.725	0.676
Garlic-In-Oil (IYH)	557	0.826	0.865	0.845	0.680	0.632	0.710	0.674
Probiotics (ERO)	377	0.876	0.818	0.846	0.674	0.659	0.748	0.694
Trans Fat (ERO)	604	0.879	0.824	0.850	0.618	0.637	0.731	0.662
Average	605	0.864	0.842	0.853	0.686	0.658	0.743	0.696

Overall, these results indicate moderate agreement, crossing the controversial 0.67 threshold for α (cf. [5]). Details on these computations are given in the remainder of this section.

[11] Except while annotating documents used for computing inter-annotator agreement, of course.

[12] Specifically, 3,024 words were annotated by both annotators, representing 10.61% of the subcorpus.

Tool support for computing these statistics for coreference annotation is not widely available: the only such resource to our knowledge was an online agreement calculator [117], but unfortunately it is no longer available. In an effort to improve the situation, the Java source for the tools used to compute the above statistics is provided at `http://www.cs.uwaterloo.ca/~dhirtle`.

3.4.3.1 Precision, Recall and F-Measure

The MUC scoring scheme assumes a gold standard exists, which is not true in our case. Instead, we follow the common practice of treating the annotations of the most experienced annotator as a gold standard, or 'key', and those of the other annotator as a 'response'. As markables are not predefined, we also incorporate a string overlap measure when computing precision and recall as done for the DAARC 2007 Anaphora Resolution Exercise [106]:

$$\text{overlap}(s_1, s_2) = \frac{\text{words}(\text{overlap}(s_1, s_2))}{\max(\text{words}(s_1), \text{words}(s_2))}$$

This overlap measure serves to reward partial agreement, e.g., when one annotator marks *the doctor* while the other marks only *doctor*. It is used to compute the amount of intersection (I) between two annotations A and B as shown in Algorithm 1. Precision, recall and F-measure are then defined as follows:

$$P = \frac{I}{|A|} \qquad R = \frac{I}{|B|} \qquad F = \frac{2PR}{(P+R)}$$

where $|A|$ and $|B|$ are the number of coreference links in the response and key, respectively.

3.4.3.2 Krippendorff's α

For our purposes, the α statistic has two main advantages over the MUC scoring scheme:

1. factoring in the likelihood of annotators (dis)agreeing by chance, and
2. incorporating a distance metric δ_{ij} to address the partial agreement issue.

Both of these advantages are made evident by its equation:

$$\alpha = 1 - \frac{\text{P(observed disagreement)}}{\text{P(expected disagreement)}} = 1 - (n-1)\frac{\displaystyle\sum_i \sum_{j>i} o_{ij} \cdot \delta_{ij}^2}{\displaystyle\sum_i \sum_{j>i} n_i \cdot n_j \cdot \delta_{ij}^2}$$

Algorithm 1 ANNOTATION-INTERSECTION

```
Require: Two annotations $A$ and $B$ of the same text
Ensure: The amount of intersection $I$
  $I \leftarrow 0$
  for all markables $m_b$ in annotation $B$ do
    $max \leftarrow 0$
    for all markables $m_a$ in annotation $A$ do
      $score \leftarrow \text{overlap}(m_a, m_b)$
      if $score > 0$ and $m_a, m_b$ are in corresponding chains then
        if $score > max$ then
          $max \leftarrow score$
        end if
      end if
    end for
    $I \leftarrow I + max$
  end for
```

The most common distance metric used for coreference is that of Passonneau [112]. We also provide results using Jaccard and Dice's metrics from information retrieval for comparison, as done by Poesio and Artstein [114]. These two distance metrics take the size of chains into account, an advantage over that of Passonneau.

Note that the range for α is $[0, 1]$, where 0 indicates chance agreement and 1 is perfect agreement (cf. precision and recall).

Chapter 4

Analyzing Health Education

Using the corpus whose creation was described in the previous chapter, we are able to investigate specific characteristics of health education that are likely to influence the design of our coreference resolution algorithm (Chapter 5). These characteristics relate to readability, pronoun distribution and, most importantly, coreference. The three sources used in this corpus (IYH, ERO and Mayo; see Section 3.1) often exhibit distinguishing features of their own. In such cases, we provide analysis for each source in addition to the corpus as a whole.

Where possible, the results of this analysis are compared to available data for previously studied domains. These comparisons show that the health education domain differs in several key aspects from, for example, news articles and technical manuals.

4.1 Readability

Readability refers to reading ease or, in other words, the level of reading skill a person must have in order to read and understand a particular text. Various tests have been devised to assess readability, with their results based on mathematical formulas that take into account surface characteristics of the text such as average sentence length. Although these readability tests are estimates that do not take all factors into account (e.g., semantic difficulty), they are accurate enough for many applications and have been widely used over the past several decades.

For the health education corpus, we use three of the most popular tests: Flesch Reading Ease [44], Flesch-Kincaid Grade Level [74] and SMOG[1] [86]. We also investigate lexical diversity using type-token ratios.

[1] SMOG is sometimes assigned the backronym "Simple Measure of Gobbledygook".

The results of these readability tests (and related statistics) are presented in Table 4.1, revealing that ERO has the highest readability and IYH the lowest. These differences in readability are probably a side-effect of covered topics: ERO provides mostly non-technical dietary and nutritional information, whereas IYH includes many subjects that require scientific explanations. For instance, disease names themselves (e.g., *tuberculosis*, *lymphogranuloma venereum* and *human papillomavirus*) are often polysyllabic.

Table 4.1: Common readability statistics for our corpus (c = characters, $\cdot$ = syllables, w = words, s = sentences)

Source	c/w	$\cdot/w$	w/s	Flesch Reading Ease	Flesch-Kincaid Grade Level	SMOG Grade
ERO	4.90	1.48	15.80	65.25	8.08	10.74
IYH	5.22	1.67	18.05	46.97	11.19	13.44
Mayo	5.20	1.66	15.80	50.56	10.13	12.21
Total	5.16	1.64	16.61	51.55	10.19	12.46

Looking at the corpus as a whole, it is clear that sampled health education articles are written at more difficult readability levels than recommended, as has often been reported in the literature (e.g., [30, 72, 75]).

4.1.1 Flesch Reading Ease

The Flesch Reading Ease formula [44], commonly used for documentation standards, results in a score (e) from 0 to 100 where higher values indicate increased readability. The formula involves the average length of sentences and words (measured in words and syllables, respectively) as follows:

$$e = 206.835 - 1.015(w/s) - 84.6(\cdot/w)$$

Table 4.2 gives average scores for random samples of several popular periodicals as computed by Flesch [46].

The Flesch Reading Ease score of 51.55 for this corpus indicates that it is considered ‘fairly difficult’ to read (10th–12th grade), bordering on the sub-50 ‘difficult’ or college-level range, according to standard score interpretation [45]. This level of readability is comparable to that of *Time* magazine, which is considerably easier to read than other news periodicals such as the *New York Times*.

As the *New York Times* is representative of news sources sampled in linguistic corpora (e.g., MUC-7 [22], the American National Corpus [122]), these readability

Table 4.2: Flesch Reading Ease scores for popular periodicals

Source	e
Reader's Digest	65
Time	52
Newsweek	50
Wall Street Journal	43
New York Times	39
Harvard Law Review	32

scores indicate that health education articles are generally easier to read than news-related ones. Intuitively, this only makes sense: news articles must pack a lot of information into small amounts of text due to space restrictions, reducing reading ease. Readability also appears to be inversely proportional to circulation. *Reader's Digest*, for instance, is aimed at a wide readership while the *Harvard Law Review* targets a much smaller scholarly audience.

4.1.2 Flesch-Kincaid Grade Level

A popular variant of the previous test is the Flesch-Kincaid Grade Level formula [74], which conveniently translates the readability score to a grade level. This facilitates its use in educational contexts, including health education where it is commonly cited (e.g., [51]). The formula for this grade level (g_{fk}) variant is quite similar to the original:

$$g_{fk} = 0.39(w/s) + 11.8(\cdot/w) - 15.59$$

As shown in Table 4.1, g_{fk} for this corpus was computed to be above 10, whereas experts recommend readability levels below grade 8 (e.g., [120, 121]). These results are consistent with other readability studies consistently demonstrating that health educational materials are not readable by the average person [49].

4.1.3 SMOG

SMOG [86] is another readability test based on grade level. It has been used in patient education for decades (e.g., [51, 144, 145]). SMOG is based on p, the number of words with three or more syllables:

$$g_{smog} = 1.0430\sqrt{p\left(\frac{30}{s}\right)} + 3.1291$$

The SMOG grade level for this corpus was found to be 12.5, which is considerably higher than the Flesch-Kincaid Grade Level (∼10). This discrepancy is due to g_{smog} representing the grade level required for *complete* comprehension [86]. A difference of two–three grades has been reported in several studies [49]. In any case, the SMOG measure further supports that the health education corpus is less readable than is recommended by health education experts [121].

4.1.4 Type-Token Ratio

Vocabulary, or lexical diversity, also influences readability. A highly varied vocabulary is likely to include words with very specific meanings, reducing ease of comprehension. One way to measure lexical diversity is using word length because short words are more common than longer ones [146]. Word length is a key component of all three readability tests described earlier.

Another approach to measuring lexical diversity in a text is **type-token ratio**, the number of unique words (types) divided by the total number of words (tokens) [135]. A ratio of 1 indicates that every word in a text occurs exactly once, whereas a low ratio indicates that many words are repeated, increasing readability. For example, the type-token ratio of academic works tends to be high, whereas that of speech tends to be relatively low [12].

Type-token ratio depends heavily on the length of a text: in a short snippet, there is likely to be little repetition, resulting in a high ratio. Conversely, a collection of many texts will include much more repetition, resulting in a low type-token ratio. For this reason, type-token ratios are often calculated for each 1,000 words of a text and then averaged.

Table 4.3: Standardized type-token ratios

Source	Ratio
News	48.27
Literature	45.02
Academic	40.76
Health Education	39.21

This standardized type-token ratio was computed for the health education corpus using WordSmith Tools [125]. Table 4.3 shows the results along with type-token ratios for academic, literary and news texts as reported by Nishina [103]. Note that the ratio for this health education corpus roughly corresponds with that of academic texts, which is considerably lower than the news type-token ratio. This ordering agrees with the analysis of Biber[2] [12].

The high readability scores and low type-token ratio for our corpus both indicate that the corpus is relatively easy to read, the former because of shorter words and sentences and the latter because of a smaller vocabulary.

4.2 Pronouns

The part-of-speech[3] breakdown given in Table 4.4 reveals that 5.2% of our corpus is made up of pronouns (approximately 3% personal and 2% reflexive). The largest contributor of pronouns is the Mayo material, where they represent 6.79% of its total words. Mayo's above-average proportion of pronouns is likely because it includes testimonials, which are usually focused on a single individual.

Proper nouns are also relevant for coreference resolution. Examining their distribution within the corpus, ERO and IYH articles appear to contain significantly more proper nouns than Mayo. However, closer examination revealed that this variation is caused by a preprocessing issue where capitalized headings are erroneously identified as proper nouns. This problem did not carry over to Mayo because headings are far less common.

As mentioned in Section 3.4.1, gerunds are sometimes treated as NPs for the purposes of coreference annotation. As this leads to extra annotation effort and potential disagreement, our annotation scheme simply follows the POS tagger. According to these statistics, gerunds and present participles are actually quite rare, comprising only 2% of the corpus, and so have little impact on overall results.

Comparing distributions of pronouns by type (Table 4.5), we see that possessive pronouns play a larger role in this corpus than they do in others. Interestingly, the overall proportion of pronouns is also higher than in other domains by as much as 4%[4].

The distribution of pronouns by grammatical person and number, given as Table 4.6, reveals that second-person pronouns constitute the majority in this corpus. First-person pronouns (especially plural ones) are abundant in ERO articles; IYH

[2] Unfortunately, Biber's type-token ratios are not directly comparable with these standardized ratios because his calculations used only the first 400 words of each text. Biber's ratios are therefore higher, e.g., 50.6 for academic, 55.3 for news.

[3] The categories correspond to the Penn Treebank tagset [82].

[4] This large gap may be a result of scope: it is unclear whether the quantity reported as 'pronouns' in [96] includes all pronouns or only third-person pronouns.

Table 4.4: Distribution of words by major parts-of-speech

Part-of-speech	IYH	ERO	Mayo	Total
Pronouns	3.48%	4.92%	6.79%	5.20%
Personal	2.11%	3.11%	4.09%	3.16%
Reflexive	0.05%	0.04%	0.07%	0.06%
Possessive	1.32%	1.77%	2.64%	1.98%
Nouns	34.29%	35.31%	30.60%	32.79%
Singular/mass	19.04%	21.31%	20.54%	20.08%
Plural	9.91%	9.93%	7.52%	8.83%
Proper, singular	5.21%	4.03%	2.51%	3.81%
Proper, plural	0.13%	0.04%	0.03%	0.07%
Adjectives	9.54%	10.44%	9.64%	9.73%
Positive	8.82%	9.19%	8.75%	8.85%
Comparative	0.51%	0.96%	0.64%	0.64%
Superlative	0.20%	0.30%	0.26%	0.24%
Adverbs	3.88%	4.04%	5.01%	4.42%
Positive	3.60%	3.65%	4.50%	4.01%
Comparative	0.19%	0.34%	0.39%	0.30%
Superlative	0.10%	0.06%	0.12%	0.10%
Verbs	16.44%	16.73%	17.44%	16.94%
Base form	4.44%	6.04%	4.55%	4.74%
Past tense	0.32%	0.53%	1.17%	0.74%
Gerund/present participle	2.14%	2.14%	1.92%	2.04%
Past participle	2.66%	1.80%	1.95%	2.20%
Present singular	2.36%	2.53%	3.09%	2.72%
to be	3.62%	3.02%	3.51%	3.48%
to have	0.90%	0.67%	1.26%	1.03%

Table 4.5: Distribution of pronouns by type

Type		Health Education	News ACE [37]	 PTB[6]	Technical Manuals [96]
All	Total	17629	7909	2063	546
	% of words	5.20%	4.21%	2.18%	1.18%
Personal	Total	10710	6084	1326	466
	% of pronouns	60.75%	76.93%	64.28%	85.35%
Possessive	Total	6726	1704	708	68
	% of pronouns	38.15%	21.55%	34.32%	12.45%
Reflexive	Total	193	119	29	11
	% of pronouns	1.09%	1.50%	1.41%	2.01%

Table 4.6: Distribution of pronouns by person and number

	First			Second	Third		
Source	Singular	Plural	All	All[7]	Singular	Plural	All
ERO	6.32%	5.75%	12.07%	56.64%	16.12%	15.17%	31.29%
IYH	0.11%	1.02%	1.13%	60.39%	18.68%	19.80%	38.48%
Mayo	7.49%	0.34%	7.83%	52.01%	32.66%	7.50%	40.16%
Total	5.38%	1.33%	6.71%	54.90%	26.52%	11.87%	38.39%

articles, on the other hand, have very few. The primary cause of this difference seems to be the frequent use of *we* in place of *you* as a rhetorical device in ERO, as in *We do not fail diets—diets fail us*[5]. Finally, Mayo articles have far more third-person singular pronouns than the other articles, yet substantially less third-person plural pronouns. This is another by-product of the individual focus of testimonials.

Taking a closer look, we see that the top 10 most frequent pronouns found in our corpus and the ACE corpus (as reported by Dimitrov et al. [37]) are very different (Table 4.7). As suggested by the prevalence of second person in this corpus, *you* and the possessive *your* are the most common pronouns, while far less common in the news corpus. Gender is another interesting divergence between these corpora: feminine pronouns are more numerous than masculine ones in the health education

[5] From "Are You A Yo-Yo?" [104].

[6] This is a 94,500 word subset of the Penn Treebank corpus that was annotated for coreference [56]. Analysis provided in [96].

[7] Singular and plural cannot be distinguished for most second person pronouns (e.g., *you*, *your* and *yours*).

corpus, whereas the opposite is true of the news corpus. This is likely because testimonials from women are more common than those from men.

Table 4.7: Top ten pronouns in health education and news

Health Education		News (ACE)	
your	27.58%	*it*	18.30%
you	26.72%	*he*	14.44%
it	13.16%	*his*	9.18%
they	5.95%	*I*	9.17%
she	4.36%	*they*	8.93%
I	3.74%	*we*	7.77%
their	3.49%	*you*	6.40%
her	3.16%	*its*	5.70%
them	2.27%	*she*	3.23%
he	2.21%	*her*	2.43%

4.3 Coreference

Coreference varies across different genres just like other linguistic phenomena. We investigate the distribution of relations (e.g., identity, synonymy) and reference types (e.g., noun phrase, appositive) among coreference chains in this corpus. As some corpora lend themselves more readily to coreference resolution than others, we also attempt to characterize the complexity of the health education corpus.

Unless otherwise specified, we only use statistics from the manually annotated subset of the health education corpus in this section.

4.3.1 Relation and Reference Type

During manual annotation (see Section 3.4), two pieces of metadata were recorded for each instance of coreference. The first was the type of relation holding between the anaphor and its antecedent; in almost all cases (over 96%), this was identity (Table 4.8). Note that this relation is by far the most widely supported among annotation schemes (e.g., MUC-7 [100]). The other relations were relatively rare, the most synonymy and generalization being present in IYH articles and Mayo having the most specialization.

Table 4.8: Distribution of coreference by type of relation

Source	Identity	Synonymy	Generalization	Specialization
ERO	97.53%	0.64%	0.91%	0.55%
IYH	94.20%	2.20%	2.65%	0.88%
Mayo	96.81%	0.18%	1.75%	1.20%
Total	96.14%	0.97%	1.82%	0.92%

Table 4.9: Distribution of coreference by type of reference

Source	NP	Appositive	Copular	Bracketed	Dashed	Cataphoric
ERO	97.53%	0.37%	0.73%	1.19%	0.09%	0.09%
IYH	95.59%	0.15%	2.57%	1.32%	0.37%	0.00%
Mayo	97.11%	1.20%	0.36%	0.84%	0.42%	0.06%
Total	96.72%	0.63%	1.19%	1.09%	0.32%	0.05%

The significance of these non-identity relations is unclear considering the difficulty of distinguishing between them, as in example (4.1).

(4.1) *medical test kits ... medical kits for self-testing ...*
home-use medical tests ... do-it-yourself medical tests[8]

The second piece of metadata accompanying each coreferential link is the type of reference (Table 4.9). The vast majority (again, over 96%) involved regular noun phrases, but noteworthy exceptions include appositives in Mayo and copula in IYH articles. Cataphora is the rarest of the annotated types.

Overall, these uncommon types of relations and coreference can be considered a low priority during resolution due to their rarity.

4.3.2 Coreferential Complexity

As suggested by Barbu and Mitkov [10], we provide statistics about the complexity of the coreference found in the health education corpus, including the average referential distance of anaphors and whether they are usually intra-sentential or inter-sentential. We also analyze the corpus for three factors that often complicate resolution: non-nominal *it*, quoted text and named entities.

[8] From “Medical Test Kits for Home Use” [17].

4.3.2.1 Referential Distance

The relative distributions of intra-sentential and inter-sentential anaphoric pronouns for this corpus and several others (as analyzed by Mitkov and Hallett [96]) are provided as Table 4.10. We find that 65% of anaphoric pronouns in this corpus are inter-sentential, roughly double the rate reported for other domains like news. The largest contributor to the high number of inter-sentential pronouns is Mayo, where the subject of the testimonial is understood and therefore only infrequently mentioned explicitly.

However, it is important to note that these rates can vary considerably even within domains of text. For example, previous work by Barbu and Mitkov [10] on a smaller version of the technical manual corpus reported in Table 4.10 yielded an intra-sentential rate of 33.43% (as opposed to 51.74%).

Table 4.10: Distribution of anaphoric pronouns (intra-sentential vs. inter-sentential)

		Intra-sentential		Inter-sentential	
Source	Pronouns	Total	%	Total	%
ERO	443	209	47.18%	234	52.82%
IYH	394	206	52.28%	188	47.72%
Mayo	968	213	22.00%	755	78.00%
Total	1805	628	34.79%	1177	65.21%
Tech. Manuals	545	263	48.26%	282	51.74%
News (PTB)	1868	1323	70.82%	545	29.18%
Narrative[9]	205	127	61.95%	78	38.05%

Table 4.11: Average distance (in sentences) between anaphors and antecedents

Source	Pronominal	Nominal	All
ERO	1.56	3.91	3.05
IYH	1.25	4.59	3.74
Mayo	2.90	7.76	5.19
Total	2.21	5.39	4.14

A more fine-grained approach is to examine actual referential distance (Table 4.11). On average, pronouns in this corpus refer more than twice as far as those in technical manuals, news and narrative texts according to the distances reported by

[9] This sample consists of the first two chapters of *From the Earth to the Moon*, an 1865 novel by Jules Verne.

Mitkov and Hallett [96] (1.06, 0.26 and 0.33 sentences, respectively). This difference is largely a consequence of second-person pronouns and the inter-sentential tendency of Mayo. Unsurprisingly, nominal anaphors generally refer much further than pronominal ones, and anaphors in Mayo reach further than those in ERO and IYH.

4.3.2.2 Non-nominal *it*

Unlike other pronouns, *it* frequently does not anaphorically refer to a noun phrase antecedent. In some cases, it takes a verb phrase or entire sentence as antecedent, or has no (explicit) antecedent at all (see Section 2.1.1). Even early research recognized this problematic phenomenon. For example, Lappin and Leass [79] found that 7.7% of pronouns in their training corpus of technical manuals were pleonastic[10].

Despite the preoccupation with pleonastic *it* found in the literature (e.g., [37, 79]), all non-nominal uses must be recognized by anaphora resolution systems to prevent the default behaviour of seeking a nominal antecedent (as pointed out in [107]). 'Non-nominal' simply means that the pronoun does not have a nominal antecedent—if it has one at all, the antecedent is something other than an explicit noun phrase (e.g., a verb, clause or implied subject).

Although *it* represents a relatively small proportion of total words (0.68% in this corpus; 0.77% [37] and 0.86% [41] in others), it accounts for 13.16% of all pronouns in the full corpus (Table 4.12). According to the analysis by Dimitrov et al. [37], *it* is even more prominent in the news domain, representing over 18% of all pronouns.

Table 4.12: Total and non-nominal *it* (full corpus)

		All *it*		Non-nominal *it*	
Source	Pronouns	Total	% of pronouns	Total	% of all *it*
ERO	2643	370	14.00%	183	49.46%
IYH	4610	728	15.79%	243	33.38%
Mayo	10376	1222	11.78%	396	32.41%
Total	17629	2320	13.16%	822	35.43%

Evans [40, 41] found that *it* was non-nominal ∼32% of the time in a general corpus consisting of politics, science, fiction and journalism texts. According to our analysis, the health education domain has a slightly above average proportion of

[10] This unusually high figure is probably due to the fact that Lappin and Leass were only considering third-person pronouns. We found that non-nominal *it* (encompassing more than just pleonastic cases) represented only 4.7% of pronouns in the health education corpus.

non-nominal *it* (35.43%). Mayo had the least *it*, non-nominal or otherwise, due to its personal focus. The highest rate, approaching 50%, was found in the ERO articles, which contain a lot of idiomatic uses such as *it's as simple as that* and *spice it up*.

Considering its non-nominal tendency in this corpus, *it* must be given special consideration during resolution. The usual approach to pleonastic *it* recognition relies on patterns. Our analysis has revealed that the following pattern is highly effective for this corpus:

> it (***modal***)* ***verb*** (***modifier***)* ***particle***

where[11]:

> ***modal*** = a modal (e.g., *can, must, would*) or auxiliary verb (e.g., *has*)
>
> ***verb*** = {*is, come, feel, mean, take, believe, become, remain, hurt, seem, appear, follow*}
>
> ***modifier*** = an adjective, adverb, coordinating conjunction or cognitive verb (e.g., *recommend, assume, believe*)
>
> ***particle*** = {*to, for, that, whether, than, if, what, how, why, which, who*}

This pattern essentially generalizes and extends the first three identified by Lappin and Leass [79]. The set of cognitive verbs included in ***modifier*** is the same as the expanded set reported in [37].

In fact, this single pattern accounts for roughly half of the non-nominal or pleonastic uses of *it* in this corpus (see Section 5.5.2 for details). By comparison, the greatest coverage achieved by any single pattern for the ACE corpus was 25.8% as reported by Dimitrov et al. [37].

4.3.2.3 Quoted text

Quoted text poses a particular problem for the resolution of certain pronouns because they can behave very differently in this context. For example, the *I* and *he* corefer in the following example:

(4.2) *"I had no pain and no bleeding from the anesthesia," he says.*[12]

[11] The * is the Kleene star, which represents zero or more repetitions as for regular expressions. Also note that morphological variation is taken into account for all parts of the pattern.

[12] From "Enlarged prostate (BPH)" [24].

This is normally only a problem for speech and dialogue, not other uses of quotes as in *“whole grain” bread* or *“good” fat.*

As shown in Table 4.13, the amount of quoted text varies significantly by source. IYH has very little, mostly single words, whereas Mayo has frequent quotes from the subject of the testimonial (e.g., about their decisions and experiences) that are generally quite long. This variation in the amount of dialogue means that special consideration by the resolution algorithm may be problematic.

Table 4.13: Number and average length of quotes (full corpus)

	Quoted words		Words/quote	
Source	Total	%	Mean	Max
ERO	1193	2.22%	8.52	62
IYH	334	0.25%	2.04	28
Mayo	8886	5.82%	15.84	165
Total	10413	3.07%	12.04	165

4.3.2.4 Named entities

Successful coreference resolution in many domains is highly dependent upon the ability to accurately identify names of people, organizations and geographic locations. In the case of health education, only people names are sufficiently common to warrant attention. First names are particularly important because the gender information that they carry is crucial during resolution.

Table 4.14 shows the number of times that first names occur in our corpus as identified by the named entity recognizer included with GATE [28]. Male and female names are approximately equal in number. The vast majority of the names, over 90% in fact, came from Mayo.

Table 4.14: Quantity and gender of name occurrences (full corpus)

Source	Female	Male	Total
ERO	11	27	38
IYH	25	29	54
Mayo	501	463	964
Total	537	519	1056

Chapter 5

Resolving Coreference in Health Education

Using the corpus and analysis discussed in the previous two chapters, we developed an algorithm for resolving coreference in health education texts. As this is a new domain for coreference analysis, our focus was not on performance of the algorithm[1] but rather on using it as a means to investigate coreference and its resolution in health education.

To this end, the algorithm's scope of resolution (explained in Section 5.1) is as large as possible, encompassing not just pronouns but noun phrases in general. The algorithm itself is not intended to be innovative—by adhering to convention, we ensure that results are at least somewhat comparable with other domains. The overall design of the algorithm has nevertheless been partially influenced by features of this domain, as outlined in Section 5.2.

Because coreference is resolved differently depending on the type of anaphor, the resolution algorithm is best viewed as two interrelated procedures: one for resolving pronoun anaphors, the other for noun anaphors[2]. Section 5.3 describes each in detail, and the algorithm's actual implementation is discussed in Section 5.4.

Despite its ambitious scope and unsophisticated design, the resolution algorithm performs surprisingly well (as detailed in Section 5.5 and discussed in Section 5.6), suggesting that the health education domain provides favourable conditions for coreference resolution.

[1] After all, there is no direct basis for comparison.

[2] For convenience, we sometimes refer in this chapter to 'noun anaphors' and 'noun resolution', which are to be understood as the counterparts to pronoun anaphors and resolution. A noun in this context is any non-pronominal NP.

5.1 Scope of Resolution

We deliberately avoided restricting the algorithm's scope of resolution with the intention of learning as much about the health education domain as possible. In particular, the algorithm resolves coreference among general noun phrases (including common nouns, proper names and personal, possessive and reflexive pronouns) and handles difficult cases such as non-nominal *it*.

However, some varieties of anaphora are still poorly understood. As we did during the manual annotation of the corpus, we disregarded these varieties and instead focused on identity-of-reference direct nominal anaphora[3], which has been most widely studied in the literature [92].

During manual annotation, there were sporadic instances of demonstrative (e.g., *this, that, these*), indefinite (e.g., *something*) and relative (e.g., *that, which*) pronouns being judged to be nominally anaphoric. These exceptional cases are considered out of scope for the algorithm.

5.2 Designing for Health Education

The basic approach of the algorithm is knowledge-lean, neither relying on complex knowledge sources (e.g., grammars, lexicons, ontologies) nor deep analysis (e.g., syntactic parsing). The nature of this approach makes it inexpensive both in terms of resources during development and computation during execution. The knowledge-lean approach is common for pronoun resolution (e.g., [6, 37, 71, 88, 94]), but is typically not enough for full coreference resolution, where rich knowledge sources and/or full syntactic parsing are used (e.g., [7, 67, 116]). For this domain, however, our resolution algorithm demonstrates that the knowledge-lean approach can be successfully applied to full coreference involving general noun phrases.

The corpus analysis from Chapter 4 provided essential details about the coreferential features of the health education domain. From this data, we made a number of observations that directly influenced the design of the resolution algorithm:

O1. Second-person pronouns alone constitute over 50% of total pronouns, making their effective resolution a high priority.

O2. Reflexive pronouns are as rare (∼1%) as in other domains (e.g., [37]) and can therefore be handled in the same manner as other pronouns without significance loss of performance, regardless of their distinctive referential properties.

[3] See the annotation scheme (Section 3.4.1) for details.

O3. The average distance between a pronoun and its antecedent in this corpus is greater than two sentences, whereas in other domains this distance is a single sentence or less [96].

O4. The amount of speech and dialogue varies considerably between sources included in the corpus (rare in IYH, but fairly common in Mayo).

O5. Names are common only in Mayo texts.

O6. A significant proportion of occurrences of *it* are non-nominal, especially in ERO articles, but these cases often follow a predictable pattern.

As pointed out in Section 4.2, O1 is a major difference between the health education domain and others, such as news. In research focusing on these other domains (e.g., [94]), second-person pronouns are rarely included. When this exclusion is explicitly mentioned at all, the most common reasons provided as justification are:

1. Second-person pronouns are mostly found in speech or dialogue, where they are difficult to resolve [138].
2. These pronouns are usually not anaphoric anyway because they are deictic, i.e., dependent on extra-linguistic context for their interpretation [9, 92].

The first reason is generally true, but speech and dialogue are much less common in health education than in other domains (cf. O4). The second reason, on the other hand, is not entirely accurate, and certainly not reason enough to entirely dismiss second-person pronouns. For example, the instances of *your* in (5.1) clearly have the same referent as *you*, regardless of whether or not the exact identity of that referent is known. Therefore, this is an instance of coreference. We can also treat *your* as an anaphor and *you* as its antecedent in much the same way as *her* and *the woman*, regardless of the exact identity of the woman.

(5.1) ***You*** *can improve* ***your*** *health by increasing* ***your*** *level of physical activity.*

Furthermore, such deixis essentially becomes a non-issue when the extra-linguistic context is known, as in domain-specific situations such as this one. In health education, for example, the unique referent of *you* (outside of dialogue, at least) is the consumer of the health information (e.g., a patient).

For these reasons, and because second-person pronouns play such an important role in this domain, the resolution algorithm handles them. Though much less common in this corpus, first-person pronouns are also included. In other words, all personal, possessive and reflexive pronouns are resolved.

While analysis by Hobbs [66] and others indicates that the current sentence and two previous ones are almost always sufficient when searching for the antecedent of a pronoun, O3 indicates that a larger number of preceding sentences is appropriate for this corpus.

According to O4, if the resolution algorithm assigns too much importance to dialogue and speech, performance may suffer on the ERO and IYH articles. Therefore a quoted text module is not included with the algorithm.

The same risk applies to name recognition (O5). However, the gender information that names provide is indispensable for gender agreement, making recognition worthwhile regardless of occasional false positives. In other domains, named entities include organizations, locations and even expressions of quantities (e.g., currency), dates and times. Our analysis of the health education corpus revealed that only person names are significant (see Section 4.3.2.4), so the algorithm does not need to handle other named entities.

The algorithm also includes a non-nominal *it* filter. In keeping with the knowledge-lean approach, however, the filter is fairly basic: it consists of a single pattern, given in Section 4.3.2.2. Fortunately, O6 suggests that even rudimentary filtering should prove beneficial.

For noun resolution, we rely on lemmatized string matching. Previous research has revealed that string matches are a highly reliable indicator of coreference. For example, Soon et al. [130] found that the string match feature of their machine learning algorithm performed almost as well as all other features combined.

5.3 The Resolution Algorithm

While coreference is a general relation that can occur between any two referring expressions, it is useful in practice to divide its resolution into subproblems based on the type of referring expression. We do so here, presenting our resolution algorithm in two parts: one for pronoun resolution and the other for noun resolution. Interestingly, humans also seem to employ such a 'divide and conquer' approach, processing some anaphors differently than others [52, 59].

5.3.1 Pronoun Resolution

Pronoun resolution is a well-studied subset of coreference resolution, one where traditional rule-based approaches generally perform as well as learning-based ones [8, 133].

We call the portion of our algorithm dedicated to resolving pronoun anaphors PRONOUN-RESOLUTION, included as Algorithm 2. In addition to the set of pronouns to be resolved (P), it also accepts a parameter s_{max} indicating how many

Algorithm 2 PRONOUN-RESOLUTION

```
   Input: Set of pronouns $P$ found in document $D$,
          maximum $s_{max}$ of preceding sentences to search for antecedents
   Output: An antecedent $A[p]$ for each anaphoric pronoun $p \in P$
 1. $prev_1 \leftarrow \varnothing$
 2. $prev_2 \leftarrow \varnothing$
 3. for all pronouns $p \in P$ do
 4.     // First- and second-person pronouns usually corefer with previous ones
 5.     if $p$ is first-person then
 6.         $A[p] \leftarrow prev_1$
 7.         $prev_1 \leftarrow p$
 8.     else if $p$ is second-person then
 9.         $A[p] \leftarrow prev_2$
10.         $prev_2 \leftarrow p$
11.     else
12.         if $p$ is anaphoric then
13.             // Get candidate antecedents (in reverse document order)
14.             $s \leftarrow$ sentence in $D$ that contains $p$
15.             $C \leftarrow$ all noun phrases that precede $p$ in sentences $[s, s - s_{max}]$
16.             $A[p] \leftarrow$ SELECT-ANTECEDENT($p$, $C$)
17.         else
18.             // $p$ is non-nominal 'it'
19.             $A[p] \leftarrow \varnothing$
20.         end if
21.     end if
22. end for
```

preceding sentences to consider when searching for an antecedent. We use $s_{max} = 5$, which was empirically derived during corpus analysis.

PRONOUN-RESOLUTION loops through all pronouns in P. First-person and second-person pronouns receive special attention (as explained in Section 5.2) because they normally refer back to one another. Maintaining a pointer to the last occurrence of these pronouns (lines 7 and 10) is an optional strategy that serves to eliminate unnecessary processing.

For third-person pronouns, a check is first made to determine if the current pronoun p is indeed anaphoric using a filter to detect instances of non-nominal *it*. A list of candidate antecedents C is then generated consisting of all noun phrases preceding p up to a maximum of s_{max} sentences back (line 15). C is ordered so that candidates nearest to p in the document precede those that are further away.

The candidate list is then used by SELECT-ANTECEDENT (Algorithm 3) to find an antecedent for p. It does this by looping through all the candidates in C and choosing the first (i.e., nearest) candidate that is compatible with p. A pronoun is considered compatible with a candidate if it agrees in number, gender and animacy[4] (see Section 2.2.1). In the rare case that no compatible candidates are found, the algorithm defaults to selecting the nearest noun phrase.

Algorithm 3 SELECT-ANTECEDENT (Pronoun)

```
Input: Pronoun $p$ and set of candidate antecedents $C$
Output: The antecedent of pronoun $p$
$fallback \leftarrow \varnothing$
for each candidate $c \in C$ do
    // Return nearest compatible candidate
    if $c$ is an anaphoric pronoun then
        if $c$ and $p$ agree in number, gender, animacy and person then
            return $c$
        end if
    else if $c$ is a noun phrase then
        $c \leftarrow$ head of its noun phrase
        if $c$ and $p$ agree in number, gender and animacy then
            return $c$
        else if $fallback = \varnothing$ then
            $fallback \leftarrow c$
        end if
    end if
end for
// Default to nearest NP if no compatible antecedent found
return $fallback$
```

5.3.2 Noun Resolution

Whereas the vast majority of pronouns are anaphoric, noun phrases often introduce new entities into the discourse instead of referring to existing ones. For example, roughly 50% to 60% of definite noun phrases are non-anaphoric [11, 141]. Distinguishing between these cases is essential for proper resolution [142].

Another major difference from pronoun resolution is that the antecedents of noun anaphors are usually found much further away [118]. In this corpus, for example, the average referential distance of noun anaphors is 5.39 sentences but only 2.21 sentences for pronouns (see Section 4.3.2). Because of this, a maximum distance as used by PRONOUN-RESOLUTION is inappropriate.

We refer to the part of our algorithm dedicated to resolving noun anaphors as NOUN-RESOLUTION (Algorithm 4). The initial step (lines 2–9) differs from that

[4] Grammatical person is also checked if the candidate happens to also be a pronoun, to prevent *he* from referring to *me*, for example.

Algorithm 4 NOUN-RESOLUTION

```
Input: Set of nouns $N$ found in document $D$
Output: An antecedent $A[n]$ for each anaphoric noun $n \in N$
// Create list of markable nouns $M$
$M \leftarrow N$
for all nouns $n \in N$ do
    if $n$ is a stop word then
        $M \leftarrow M \setminus n$
    else if $n$ is only a prenominal modifier in $D$ then
        $M \leftarrow M \setminus n$
    end if
end for
// Find coreference between markable nouns only
for all markable nouns $m \in M$ do
    $l \leftarrow$ lemma of $m$
    // Get candidate antecedents (in reverse document order)
    $C \leftarrow$ all markable nouns preceding $m$ in $M$ that match $l$
    $A[m] \leftarrow$ SELECT-ANTECEDENT($m$, $C$)
end for
```

Algorithm 5 SELECT-ANTECEDENT (Noun)

```
Input: Markable noun $m$ and set of candidate antecedents $C$
Output: An antecedent of noun $m$, if applicable
for each candidate $c \in C$ do
    // Return nearest compatible candidate
    if $c$ is the head of a noun phrase then
        $s \leftarrow$ noun phrase that contains $c$ (minus initial articles)
    else // $c$ is a prenominal modifier
        $s \leftarrow$ lemma of $c$
    end if
    if $c = s$ then
        return $c$
    end if
end for
// No compatible antecedents, so consider it to be non-anaphoric
return $\varnothing$
```

of PRONOUN-RESOLUTION in that a list of relevant markables M is first created by discarding nouns that are stop words (e.g., *fun, problem, example*) or that only occur as prenominal modifiers in the text (e.g., *health* or *care* in *health care worker*). The strategy is then simple: loop through each markable m in M and create a list of candidate antecedents C consisting of all preceding markables that match the lemma of m. Using lemmas permits some morphological variation; the lemmas of *child* and *children* match, for example.

The candidate list C is then used in the noun version of SELECT-ANTECEDENT, presented as Algorithm 5. Much like the version used by PRONOUN-RESOLUTION, the nearest compatible candidate is selected as the antecedent of m. Instead of using grammatical agreement, however, compatibility is based on lemmatized string comparison, excluding articles such as *a, an* and *the*. Because nouns are often non-anaphoric, as mentioned earlier, no fallback is chosen should a compatible candidate not be found.

5.4 Implementation

The algorithm is implemented in Java and uses only free or open source third-party resources. The full source is available at `http://www.cs.uwaterloo.ca/~dhirtle`.

As the corpus is encoded in XML (see Section 3.2), resolution involves significant XML processing. A tree-based API called XOM[5] greatly simplifies this aspect of resolution, especially where XSLT and XPath are involved.

Preprocessing is a crucial part of any coreference resolution system. Unfortunately, errors are inevitable during this stage, often greatly degrading performance by providing inaccurate information to the resolution algorithm. Comparisons of manual and automatic preprocessing have shown performance differences of up to 25% [94].

For this system, preprocessing is performed by modules from GATE [28]. As discussed in Section 3.3, these modules add token, sentence and part-of-speech annotations to the text. The algorithm requires additional preprocessing in the form of noun phrase and person identification. An implementation of the Ramshaw and Marcus base NP chunker [119] outputs the former, while a named entity recognizer using a list of first names provides the latter.

[5] See `http://www.xom.nu`.

5.5 Evaluation

Following the recent trend among resolution systems, and to best represent the resolution task in this domain, we refrain from manually editing the results of preprocessing (as was done, for example, in [3, 71, 79, 88]). In other words, the resolution system is 'fully automatic'—we evaluate on real-world, not hand-crafted, input.

This section begins by detailing our evaluation methodology, including an overview of measures used. The resolution algorithm is then independently evaluated on pronouns and nouns, followed by a combined evaluation.

5.5.1 Methodology

The manually annotated portion of the health education corpus was divided into two sets: one for development and one for final testing (Table 5.1). By doing so, we avoided biasing our algorithm with the data used during its development, resulting in a more meaningful evaluation based on unseen documents. Following the example of Lappin and Leass [79], we divided the available data almost equally, with slightly more than half for the development set.

Table 5.1: Contents of the development and test sets

	Words		Pronouns	
	Total	%	Total	%
Development	15358	53.88%	928	52.58%
Test	13147	46.12%	837	47.42%

Although standard for machine learning algorithms, this methodology seems less common among rule-based systems. For example, Lappin and Leass [79] divided their data and conducted a blind test, but Kennedy and Bougaraev [71], Mitkov et al. [94] and Dimitrov et al. [37] apparently did not. This may be due to rule-based systems not requiring a formal training phase, or because error rates are more accurate and thus usually higher when evaluating them on unseen data [69, 80].

The amount of data used in this evaluation (13,147 words, 837 pronouns) is above average relative to many well-known resolution systems. For example, Lappin and Leass [79], Kennedy and Boguraev [71], Baldwin [6] and Dimitrov et al. [37] evaluated their pronoun resolution systems on less than 400 pronouns. In a recent comparative evaluation of several systems, Mitkov and Hallett [96] used data consisting of 653 pronouns. The MUC-6 and MUC-7 test corpora used for the coreference task consisted of about 13,400 and 10,000 words, respectively.

For full coreference resolution, we used precision, recall and F-measure as in MUC-7 [64, 143]:

$$P = \frac{C}{I} \qquad R = \frac{C}{T_c} \qquad F = \frac{2PR}{(P+R)}$$

where:

$C =$ the number of correctly identified coreference links,

$I =$ the number of identified coreference links and

$T_c =$ the actual total number of coreference links[6].

Note that we did not incorporate a string overlap measure as in Section 3.4.3.1 because doing so would have reduced compatibility with the results reported by other systems.

For evaluating pronoun resolution, however, we used a different measure because precision and recall are always equal when resolving all anaphors. Furthermore, as noted by Mitkov [89] and Byron [16], the computation of recall could either be based on the number of anaphors identified by the algorithm (as in [3]) or on all annotator-identified anaphors (as in [6]).

For these reasons, Mitkov [89] proposed the **success rate** measure, defined as:

$$\text{Success rate} = \frac{\text{Number of successfully resolved anaphors}}{\text{Number of all anaphors}}$$

Success rate itself is imperfect, unfortunately, in that correct filtering of non-nominal *it* does nothing to increase it. In fact, attempts to filter non-nominal *it* can decrease success rate, i.e., by incorrectly filtering cases of *it* that are actually anaphoric. Therefore we also included **resolution etiquette**, a measure later introduced by Mitkov et al. [94] that rewards the correct identification of non-nominal *it*. It is defined as follows:

$$\text{Resolution etiquette} = \frac{A+N}{T_p}$$

[6] As identified during manual annotation.

where:

A = the number of correctly resolved anaphoric pronouns,

N = the number of correctly filtered non-anaphoric pronouns and

T_p = the total number of pronouns (anaphoric or otherwise).

As done for MUC-7 scoring [100], we consider resolution to be successful if at least the head of the noun phrase corresponding to the correct antecedent is identified. This flexibility is important because errors during preprocessing may include incorrect noun phrase boundaries. For example, the NP chunker might detect the boundaries of the noun phrase *the surgery last week* as *[the surgery] [last week]*. If the algorithm identified only *the surgery* as the antecedent of a later reference to this surgery, this would still be scored as correct because *surgery* is the head of the noun phrase.

Some evaluation methodologies, such as that used in [96], do not prevent the scenario where certain pronoun resolution errors may cause additional ones. For example, if a pronoun a refers to another pronoun b, but b has been resolved incorrectly, a would also be considered incorrect. In our evaluation, we prevent errors from chaining in this way (as also done, for example, in [6, 76, 136]). In fact, we have no choice but to adopt this approach because initial first-person and second-person pronouns often never have a textual antecedent.

5.5.2 Pronoun Resolution

We first evaluated the performance of the resolution algorithm on personal, possessive and reflexive pronouns. As shown in Table 5.2, PRONOUN-RESOLUTION (Algorithm 2) successfully resolves 86.65% of pronouns in the test set[7]. The IYH texts were the most challenging in terms of resolution by a margin of 7–8%. Conversely, the Mayo part of the corpus was the easiest to resolve by about 17%.

Of the three types of grammatical agreement enforced by the algorithm, number agreement had the greatest individual impact on the overall score, and was especially important for the ERO and IYH parts of the corpus. Disabling *both* gender and animacy agreement had a substantial effect, but only for Mayo.

During development of the algorithm, the optimal window for antecedent selection was determined to be five sentences. For the test set, it turned out that using six sentences would have narrowly improved the success rate. A complete analysis of

[7] The success rate achieved on the development set was only slightly higher (87.4%).

Table 5.2: Success rate of pronoun resolution by types of grammatical agreement

		Disabled agreement			
Source	Default	Number	Gender	Animacy	Gender+Animacy
ERO	83.23%	76.65%	83.23%	83.23%	83.23%
IYH	75.56%	66.67%	75.56%	75.56%	75.98%
Mayo	92.59%	88.89%	91.20%	92.36%	53.47%
Total	86.65%	81.13%	85.88%	86.52%	65.04%

Table 5.3: Success rate of pronoun resolution by search window size. The window includes the sentence containing the pronoun plus n previous sentences (default 5).

Source	0	1	2	3	4	5	6+
ERO	78.44%	83.83%	83.83%	83.23%	83.23%	83.23%	83.23%
IYH	72.22%	75.56%	75.56%	75.56%	75.56%	75.56%	75.56%
Mayo	53.47%	80.09%	87.04%	90.97%	92.13%	92.59%	92.82%
Total	63.16%	79.85%	83.70%	85.75%	86.39%	86.65%	86.78%

Table 5.4: Success rate of pronoun resolution by pronoun type

Source	Personal	Possessive	Reflexive	All
ERO	77.78%	89.83%	-	83.23%
IYH	70.25%	94.12%	50.00%	75.56%
Mayo	90.38%	97.08%	100.00%	92.59%
Total	83.17%	94.74%	80.00%	86.65%

how success rate varies according to window size is given as Table 5.3. Interestingly, performance levels peaked at a window size of one sentence for ERO and IYH articles, yet they steadily increased for Mayo until reaching a plateau at six sentences. Conversely, Mayo fared much (∼20%) worse than the others at the minimum window size.

With a success rate approaching 95%, the algorithm was most effective at resolving possessive pronouns (Table 5.4). The success rate for reflexives, on the other hand, was slightly below average. Consistent with earlier observations, however, the test set contains only a handful of reflexives; many more would be required to form an accurate picture of how well they are handled by the algorithm.

On the development set, the *it* filter was able to correctly filter 33/59 (55.93%) instances of non-nominal *it*. As shown in Table 5.5, the *it* filter was substantially less effective on the test set. This may be a result of the test set having less to work with, containing only 38 instances of non-nominal *it* compared to 59 in the development set.

Table 5.5: Accuracy of non-nominal *it* filter

		Identified cases	
Source	Cases	Total	%
ERO	16	6	37.50%
IYH	9	7	77.78%
Mayo	13	2	15.38%
Total	38	15	39.47%

The effectiveness of the *it* filter is respectable considering its simplicity, consisting of only a single pattern. It was able, for example, to identify the following cases[8] of non-nominal *it*:

(5.2) ***It** is estimated that 90% of children who are not vaccinated for chickenpox will get it by the time they are twelve.*

(5.3) *If you are thirsty or have a dry mouth, **it** is likely that you are not drinking enough water.*

(5.4) *"**It**'s just nice to get out in the boat."*

(5.5) *If you eat in front of the TV pay attention to the kinds and amounts of foods you eat as **it**'s easy to lose track and have more than you need.*

[8] From "Chickenpox Vaccine" [17], "Facts on Fluids - How to stay hydrated" [104], "Carpal tunnel syndrome" [24], "Managing Family Meals" [104] and "Halloween Safety" [17], respectively.

(5.6) *However **it** is important to keep in mind all aspects of your child's safety when planning costumes.*

Fortunately, the *it* filter did not degrade the success rate at all[9]. The resolution etiquette measure, shown in Table 5.6, revealed that the filter did make a positive contribution, albeit a small one. The largest improvement (3.27%) was seen for ERO articles, where non-nominal *it* is most common (see Section 4.3.2.2).

Table 5.6: Resolution etiquette

Source	Default	No *it* filter
ERO	79.23%	75.96%
IYH	74.60%	71.96%
Mayo	90.34%	89.89%
Total	84.21%	82.62%

5.5.3 Noun Resolution

We also evaluated the performance of the algorithm on common nouns and proper names, i.e., non-pronominal noun phrases. Establishing coreference among nouns is a much greater challenge than resolving pronouns. For example, the approach of Strube et al. [132] achieved an F-measure over 80% for most pronouns but only 33.94% for definite noun phrases. It is therefore not surprising that our scores decreased for noun resolution.

As shown in Table 5.7, NOUN-RESOLUTION (Algorithm 4) achieved 61.28% precision and 55.42% recall (58.20% F-measure) for nouns in the test set[10]. With scores in the low 50s, ERO proved to be the greatest challenge for the algorithm, whereas scores were consistently the highest (low 60s) on the Mayo texts.

The overall performance differed only slightly when disabling the list of stop words (line 4 of NOUN-RESOLUTION) or including articles such as *a* and *the* in string comparisons (line 4 of SELECT-ANTECEDENT). Keeping articles actually resulted in the highest precision (nearly 65%), though recall diminished proportionately. Filtering prenominal modifiers had the greatest impact of the evaluated options, but the overall contribution was still quite small (2.69%).

[9] That is, success rate does not decline on the test set. The same is not true of the development set, though the drop was fractional.

[10] On the development set, the results were noticeably better: 66.94% precision, 63.57% recall (65.21% F-measure).

Table 5.7: Precision, recall and F-measure scores for noun resolution by optional features

	Default			Allow stop words			Compare even articles			Include prenominal modifiers		
Source	P	R	F	P	R	F	P	R	F	P	R	F
ERO	52.70%	50.60%	51.63%	48.47%	50.60%	49.51%	55.35%	47.41%	51.07%	46.01%	50.60%	48.20%
IYH	61.06%	56.64%	58.77%	58.51%	58.51%	58.51%	63.79%	51.75%	57.14%	54.83%	56.88%	55.84%
Mayo	68.58%	57.34%	62.46%	65.50%	57.91%	61.47%	74.17%	56.78%	64.32%	64.86%	57.34%	60.87%
Total	61.28%	55.42%	58.20%	58.07%	56.38%	57.21%	64.99%	52.42%	58.03%	55.51%	55.51%	55.51%

Table 5.8: Precision, recall and F-measure scores for noun resolution by grammatical number

	Singular			Plural			All		
Source	P	R	F	P	R	F	P	R	F
ERO	57.23%	54.49%	55.83%	51.79%	34.52%	41.43%	52.70%	50.60%	51.63%
IYH	65.31%	57.84%	61.35%	56.12%	44.72%	49.77%	61.06%	56.64%	58.77%
Mayo	71.54%	63.51%	67.29%	55.56%	28.99%	38.10%	68.58%	57.34%	62.46%
Total	65.74%	59.23%	62.32%	54.74%	37.68%	44.64%	61.28%	55.42%	58.20%

Earlier versions of the algorithm prevented substrings of named entities from being involved in coreference because they are not considered markable according to our annotation scheme (see Section 3.4.1). However, the named entity check was found during development to slightly *decrease* performance, so it was disabled. The failure of this feature is probably due to spurious proper names being identified during preprocessing, such as in capitalized headings.

Table 5.8 shows the results broken down by grammatical number of the anaphor. Plural nouns were definitely most challenging: the F-measure achieved by the algorithm on singular nouns was 17.68% higher than on plural nouns. Where the plural score really suffered was recall, particularly on Mayo articles, where it dropped to just below 29%. This is a rare situation in that scores for Mayo were normally the highest.

5.5.4 Coreference Resolution

Finally, the algorithm was evaluated based on how well it resolved coreference in general, i.e., among all NPs whether pronouns or nouns. Table 5.9 shows that, as one might expect, the results lie between those for pronoun and noun resolution (86.65% and 58.20%, respectively): 72.81% precision and 68.84% recall (70.77% F-measure). As before, the resolution algorithm performed best on Mayo, the margin being greater than 15% in this case.

Table 5.9: Precision, recall and F-measure scores for coreference resolution

Source	P	R	F
ERO	65.20%	63.64%	64.41%
IYH	65.57%	62.23%	63.86%
Mayo	82.83%	76.72%	79.66%
Total	72.81%	68.84%	70.77%

5.6 Discussion

The evaluation results from the previous section confirm that the coreference resolution algorithm is remarkably effective for this domain. In particular, the 86.65% accuracy achieved for pronoun resolution disproves the hypothesis that semantic knowledge is required to reach success rates greater than 75% (cf. [93, 109]).

As already mentioned, our end objective was to investigate how coreference resolution differs for health education relative to other domains. To provide context for the results of our evaluation, there needs to be some basis for comparison. Unfortunately, as this is a new domain, scores reported for existing algorithms are not directly compatible with our own. However, recent evaluations done by Mitkov and Hallett [96] suggest that algorithm performance does not drastically differ across genres, even when the algorithm in question was originally developed on data from a specific domain. In light of this, we compare our results with those of several prominent pronoun and coreference resolution systems.

According to Mitkov and Hallett's findings (reproduced as Table 5.10), existing pronoun resolution systems achieve success rates up to approximately 60% when operating in a fully automatic manner. On the health education corpus, our knowledge-lean approach achieved scores higher than the best of these by over 25%, including those that benefit from full syntactic parsing. While unlikely on other domains, this level of performance is unparalleled[11] among fully automatic approaches for resolving personal, possessive and reflexive pronouns.

Table 5.10: Success rates of popular pronoun resolution systems as evaluated on different corpora by Mitkov and Hallett [96]

System	Tech. Manuals	News	Literature	Overall
L&L [79]	60.37%	61.40%	53.80%	60.65%
Hobbs [66]	57.98%	61.29%	53.80%	60.07%
Mitkov [88]	48.99%	59.58%	52.72%	57.03%
K&B [71]	53.58%	53.26%	48.36%	52.08%
Baldwin [6]	33.94%	39.02%	34.78%	37.66%

For general coreference resolution, the rule-based resolution systems that participated in later MUC competitions are the most suitable for comparison with our algorithm. FASTUS [4, 70] achieved the highest scores during MUC-6 [134] with an F-measure[12] of 64.85%. For MUC-7 [21], the leader was LaSIE-II [67] at 61.8%. More recently, the learning-based approach of Soon et al. [130] achieved F-measures of 62.6% and 60.4% on the same data (MUC-6 and MUC-7, respectively). Although some of these systems incorporated knowledge-rich features such as full parsing, especially LaSIE-II, our health education resolution algorithm still achieved a superior F-measure (70.77%) within its domain.

The success of our straightforward approach indicates that coreference resolution may simply be more effective for health education texts than those from other

[11] Note that other systems may report higher scores, but those that do were either not automatic or handled only a limited set of pronouns. For example, Mitkov's robust approach [88] is reported as achieving a 89.7% success rate but all preprocessing errors were manually corrected, non-anaphoric pronouns were removed and only third-person pronouns were handled.

[12] F-measures were not officially used in MUC-6 (unlike MUC-7), but are provided here for consistency.

domains. As discussed in Section 4.1, the overall readability of this corpus is significantly higher than that of news sources (such as *New York Times*) included in existing linguistic corpora. This strongly implies a link between resolution ease and readability, though density of named entities could also be a significant factor.

It is important to note that a number of improvements could be made to the algorithm and indeed might have been if performance were our principle aim:

- Pronoun resolution
  - Add additional factors for resolution (see Section 2.2), both constraints (e.g., binding theory [79]) and preferences (e.g., repetition, syntactic role).
  - Extend the non-nominal *it* filter with additional patterns, e.g., by adapting those used in [37, 79].
- Coreference resolution
  - Add detection and special processing for appositives, copula and bracketed/dashed constructions.
  - Include support for direct relations beyond identity, i.e. synonymy, generalization and specialization. This would require a semantic resource such as WordNet [43], whose inclusion would mean straying from the knowledge-lean approach.

Performance might best be improved by focusing on preprocessing. Error analysis revealed that the majority of errors actually stem from imperfect preprocessing across all levels:

1. Tokenizer
   - Dashes are treated as token boundaries where inappropriate:

     ```
     <w>Omega-</w><w>3</w> <w>bread</w>
     <w>B-</w><w>vitamins</w>
     <w>Garlic-in-</w><w>oil</w>
     ```

   - Otherwise, a boundary is appropriate but the dash should be an independent token. For example, *Here's how—these everyday classics will give your mealtimes more colour, crunch and nutritional punch*[13] is tokenized as:

     ```
     <w>Here</w><w>'s</w> <w>how-</w><w>these</w>...
     ```

[13] From "Meal makeovers that pack more veggies and fruit" [104].

2. Sentence splitter

   There can be large gaps where no sentences are detected, caused by the lack of final punctuation in headings and bulleted lists as in this example[14]:

   ```
   <s>
     ... lowering your risks for:
       * Heart disease
       * Osteoporosis (brittle bones)
       * Type 2 diabetes
     So, what is holding you back?
   </s>
   ```

   This was a significant problem for this corpus because of the number of bulleted lists, sometimes even stranding pronouns and potential antecedents. To address this issue, an XSLT stylesheet was incorporated to insert sentence boundaries before particular HTML end tags such as `</h1>` and `</ul>`.

3. Part-of-speech tagger

   Words in title case are regularly identified as proper nouns. For example, *Facts on Fluids* (the title of an ERO article) is tagged as:

   ```
   <h2>
     <w category="NP">Facts</w>
     <w category="IN">on</w>
     <w category="NP">Fluids</w>
   </h2>
   ```

   instead of:

   ```
   <h2>
     <w category="NNS">Facts</w>
     <w category="IN">on</w>
     <w category="NNS">Fluids</w>
   </h2>
   ```

   Because this error is so common, we cannot rely on words tagged as proper nouns actually being named entities.

[14] From “Get into the swing of being active” [104].

4. NP chunker

   - Base noun phrases are often not sufficient. For example, *a bowl of fresh fruit* is chunked as:

     ```
     <NounChunk><w>a</w> <w>bowl</w></NounChunk>
     <w>of</w>
     <NounChunk><w>fresh</w> <w>fruit</w></NounChunk>
     ```

     causing the first noun chunk to seem to corefer with that of *a bowl of cereal*, though it clearly does not. This causes a large number of false positives during string matching.

   - Coordinated noun phrases are inconsistently chunked:

     ```
     <NounChunk><w>fatigue</w></NounChunk>
     <w>and</w>
     <NounChunk><w>weakness</w></NounChunk>
     ```

     but:

     ```
     <NounChunk>
       <w>fruits</w> <w>and</w> <w>vegetables</w>
     </NounChunk>
     ```

     Ideally, both the coordinated noun phrase and nested noun phrases would always be identified, supporting coreference at either level.

   - Prenominal modifiers are also handled inconsistently:

     ```
     <NounChunk>
       <w>increased</w> <w>risk</w>
     </NounChunk>
     ```

     as opposed to:

     ```
     <w>carbonated</w>
     <NounChunk><w>beverage</w></NounChunk>
     ```

     where *beverage* is thereafter treated as a bare noun.

5. Person recognizer

   Occasionally last names interfere, e.g., a female having a family name identified as male:

   ```
   <Person gender="female"><w>Ann</w></Person>
   <Person gender="male"><w>Smith</w></Person>
   ```

Ideally, included preprocessing modules would be optimized for the health education domain, but of course this is not the case: TreeTagger and the NP chunker, for example, were trained on the Penn Treebank [82], which consists mostly of news.

Given this, it is quite possible that preprocessing introduced an above-average number of errors on this corpus.

One major area of difficulty is prenominal modifiers. Supporting them is necessary for recall but doing so also significantly reduces precision. For example, consider *test* and *the tests* in the following example:

(5.7) *Some of the benefits associated with home-use **test** kits are obvious. ...*
*They offer privacy because **the tests** are conducted at home.*[15]

Here *the tests* refers to the kits in general, but string matching only identifies the prenominal modifier *test* as the antecedent. Because *kits* is the head of that noun phrase, the scoring scheme treats this as incorrect[16].

In other cases, the algorithm suggested arguably valid instances of coreference that were simply overlooked or else judged to be too general by human annotators (e.g., *food, research*).

[15] From "Medical Test Kits for Home Use" [17].

[16] Unfortunately, this results in roughly 20 errors for the document because test kits are mentioned so frequently.

Chapter 6

Conclusion

This thesis has investigated coreference and its resolution in the new domain of health education. This final chapter summarizes research contributions and outlines several possible directions for future work.

6.1 Contributions

The work embodied by this thesis contributes to existing coreference resolution research in three primary ways, as elaborated in the following section.

Indirectly, this thesis also contributes to the field of health education in that coreference is an essential element of cohesive, readable text. In other words, easy-to-understand health education materials require the effective use of coreference, and this is the only known work to date exploring coreference in this domain. According to our results, health education is especially well-suited to automatic coreference resolution and we foresee its integration into software tools aiming to improve the quality of health education materials.

6.1.1 A Corpus of Health Education

We constructed a representative corpus of reliable, consumer-oriented health education texts. This corpus, which we believe to be unique, consists of 339,027 words across 230 samples taken from objective sources (Health Canada, the Government of Ontario and the Mayo Clinic). These samples cover a broad range of topics (e.g., diseases, treatments, lifestyle, nutrition) and communicative goals (information, persuasion and guidance) commonly found in health education.

The corpus was encoded in a standardized XML format and automatically annotated with relevant syntactic information such as parts-of-speech. More importantly, full coreference chains were manually annotated, using a well-defined annotation scheme and procedure, for a 10% sample of the overall corpus. Manual annotation of coreference is especially challenging because it requires complete concentration over substantial periods of time; in our case, annotating 4,120 instances of coreference across 39 texts took approximately 60 hours. Our coreference annotations were determined to be reliable according to inter-annotator agreement.

Given the lack of coreferentially annotated corpora, we plan to make this corpus, including the manually annotated portions, freely available to the wider research community for future use. We have also made our annotation guide and program for computing inter-annotator agreement publicly available (Appendix B and `http://www.cs.uwaterloo.ca/~dhirtle`, respectively).

6.1.2 In-Depth Analysis of a New Domain

To date, the vast majority of coreference resolution research has been restricted to only two domains: news and technical manuals. Health education is substantially different according to our comprehensive analysis of the constructed corpus. The distinguishing characteristics of health education relative to these other domains can be summarized as follows:

- Less complex words, sentences and vocabularies (i.e., higher readability);
- A greater overall proportion of pronouns, especially second-person ones;
- A higher ratio of inter-sentential pronouns;
- Pronouns that refer twice as far, on average;
- More frequent, but also more predictable, non-nominal *it*;
- Fewer relevant named entities (mostly person names).

In addition to these specific findings about health education, this analysis serves to expand our understanding of the extent to which coreference varies across different domains. In the long run, data about additional domains makes the ultimate goal of successful domain-independent resolution that much more attainable. If the approaches we develop are biased toward news, for example, how can we hope to achieve good results on unrestricted text?

6.1.3 A Simple but Effective Coreference Resolution Algorithm

The final contribution of this thesis is an algorithm designed for resolving coreference in health education. It resolves practically all nominal anaphora, including common nouns, proper nouns and personal, possessive and reflexive pronouns. Despite this broad scope and its knowledge-lean design, the resolution algorithm is highly effective for this domain.

The algorithm was evaluated in fully automatic mode on unseen data. This evaluation included performance on specific subsets of the input and the relative contribution of various features. The algorithm correctly resolved 86.65% of pronouns in the test corpus, disproving the hypothesis that success rates above 75% can only be attained using semantic knowledge. For coreference resolution in general, the algorithm's performance was also impressive with an F-measure of 70.77%. Error analysis indicated that there is plenty of room for improvement, especially in terms of preprocessing.

The source code of the algorithm's implementation is freely available from `http://www.cs.uwaterloo.ca/~dhirtle`.

6.2 Directions for Future Work

As this is initial research into a new domain, there are many potential directions for future work. The following four stand out in particular:

1. **Evaluate existing algorithms on this corpus**

   Our investigation indicates that coreference resolution is highly effective for health education, so existing resolution approaches should also perform well. On the other hand, it is possible that certain features of the corpus, such as generic coreference and second-person pronouns, may cause other approaches to stumble.

   More generally, are other forms of natural language processing as effective on health education text? Summarization and machine translation in particular could be very useful.

2. **Improve the coreference resolution algorithm**

   The developed resolution algorithm is fairly basic and already performs well. Would the improvements outlined in Section 5.6 significantly affect performance? It may be that these 'advanced' techniques are actually wasted on this corpus, possibly even reducing performance.

3. **Explore applications of coreference resolution in health education**

   That coreference resolution works so well on this domain encourages putting it to good use.

   For instance, the HealthDoc Project (see [35] for an overview) is developing natural language generation tools for delivering personalized health education materials. An early version of the authoring tool [110] (for creating such tailored materials) required authors to manually annotate instances of coreference, which are needed for automated personalization. Newer versions of the tool currently being developed [36] could incorporate automatic coreference resolution to remove this requirement; for cases where the resolution is incorrect, the author would merely have to make corrections instead of starting from scratch.

4. **Investigate the link between reading and resolution ease**

   Does increased readability always entail better coreference resolution? Is density of named entities also a factor? Existing research into 'resolution complexity' [10, 96] indicates that greater average referential distances imply more difficult resolution, yet this corpus seems to contradict this heuristic.

   Furthermore, can readability be used as a general indicator of suitability for NLP?

Appendix A

Health Education Corpus Samples

Special care was taken to preserve the structure and formatting of these samples, even inconsistencies and typos. As done when including the texts into the corpus, everything but the body of the text (e.g., tables of contents, glossaries, footers) has been removed.

A.1 Sample Article from EatRight Ontario

Are You A Yo-Yo?[1]

Anyone who has ever successfully lost weight knows how difficult it can be to keep it off. For most people, dropping a few pounds is just a prelude to adding a few back on—-sometimes more than they lost in the first place. Repeatedly losing and regaining weight is referred to as yo-yo dieting.

Downfalls of Yo-Yo Dieting

Yo-yo dieting is the result of low calorie diets. Restricting your eating through a low calorie diet slows down your metabolism, which causes your body to become more efficient at storing fat. Low calorie diets also lead to people ignoring their natural hunger cues and feelings of fullness; instead they rely solely on measured portion sizes and diet plans. As a result many people feel deprived which can lead to binging, weight gain, and subsequent frustration and lowered self-esteem. In an effort to lose the regained weight, another diet is attempted, and the yo-yo cycle continues. We do not fail diets- diets fail us. Therefore it is important to take a sensible approach to weight loss, to break the yo-yo cycle.

[1] © Queen's Printer for Ontario, 2007. Reproduced with permission. Available online at http://www.eatrightontario.ca/en/ViewDocument.aspx?id=62

How To Lose Weight - and Keep It Off

The overwhelming evidence is that weight loss among the overweight or obese improves various aspects of health such as risk for diabetes, coronary disease and perhaps cancer. Experts say it is still recommended that overweight and obese people try to lose weight but preferably avoid weight regain.

People who want to lose weight and keep it off should take a sensible approach that combines positive dietary changes with regular exercise. Leading a healthy lifestyle doesn't have to be difficult. By making gradual, simple changes in your eating patterns, as well as increasing your level of physical activity, you can improve your health and well-being.

When it comes to exercise for weight loss and maintenance, 60 minutes of daily aerobic exercise, such as brisk walking, is optimal, but any amount is better than none. If you are just starting out, slowly increase your level of activity over time, and try out different activities until you find something that you enjoy- from joining a sports team or taking up aerobics at the gym, to taking the dog out for a long walk or dancing around the living room for fun. And remember that it is the total amount of activity you engage in throughout the day that counts- it does not have to occur all at once.

Healthy Weight Loss Tips

- Follow Eating Well with Canada's Food Guide.
- Aim to lose one to two pounds a week- to be successful, weight loss must be gradual.
- Take smaller food portions and use smaller plates- you are less likely to over eat. You can always go for seconds if you are still hungry.
- The next time you are grocery shopping, stock up on healthy snacks such as cut up vegetables and high fiber fruits. If they are available, you are more likely to choose them when you feel the urge.
- Sit down together for a family meal, away from the television. Consider a family walk afterwards.
- Eat healthy snacks in-between meals to keep you from becoming too hungry at meal time.
- Fill 1/2 of your plate with vegetables, 1/4 with a lean protein such as a lean meat, and 1/4 of your plate with whole grains such as brown rice, whole grain pasta.

- Choose whole grain foods more often, such as whole grain bread versus white bread, or oatmeal for breakfast versus highly refined cereals. They are more nutritious, and can keep you feeling full for a longer period of time.
- Speaking of breakfast, make sure that you eat it everyday. Starting your day off with a nutritious breakfast will mean that you are less likely to overeat the rest of the day.
- When shopping for whole grains, check out the ingredient list: whole grain foods will list a whole grain - such as wheat, oats, corn or rice - as the first ingredient. Look for words "whole" or "whole grain" before the name of the grain.
- Drink water more often, and sugar-laden soft drinks or juice drinks less often- or cut them out all together.

Find out more:

- Try EATracker to learn more about your eating and activity habits
- Eating Well with Canada's Food Guide, by Health Canada
- Canada's Physical Activity Guide, by the Public Health Agency of Canada

How to lose weight without going on a diet, by Dietitians of Canada

A.2 Sample Article from It's Your Health

Chickenpox Vaccine[2]

The Issue

There are new recommendations about who should get the chickenpox vaccine. Talk to your doctor or health care provider about the benefits of this vaccine for members of your family who are at least 12 months old, and have never had chickenpox.

Background

Chickenpox (varicella) is caused by a virus called Varicella-zoster. It starts with a fever and is followed by a rash of red spots that may be itchy. There may be hundreds of these spots, which eventually turn into blisters filled with fluid. After four or five days, the blisters dry out and become crusted. From start to finish, chickenpox may last seven to ten days.

The virus spreads easily and quickly through personal contact such as touching the blisters. People with chickenpox can also spread the virus through the air when they cough or sneeze. A pregnant woman can pass the chickenpox virus on to her baby before it is born.

Most adults today who grew up in Canada had chickenpox as children. It is estimated that 90% of children who are not vaccinated for chickenpox will get it by the time they are twelve. As a general rule, you can only get chickenpox once, but it's also possible for the virus to remain in your body and become active again later on. When this happens, the virus causes a painful rash of blisters called shingles.

Complications Associated with Chickenpox

Most children who get chickenpox recover completely. However, severe cases of chickenpox can pose serious health risks, especially for newborn babies, adults, or anyone with a weakened immune system.

The complications from chickenpox can include bacterial skin infections, scars (if the blisters get infected), pneumonia, and encephalitis (inflammation of the brain). There is an increased risk of birth defects for babies who get chickenpox from their

[2] © Her Majesty the Queen in Right of Canada, represented by the Minister of Health, 2004. Reproduced with permission. Available online at `http://www.hc-sc.gc.ca/hl-vs/iyh-vsv/med/chickenpox-varicelle-eng.php`

mothers before birth. Also, children with chickenpox have an increased risk of getting necrotizing fasciatus/mytositis (flesh-eating disease). It should be noted, however, that while flesh-eating disease is a complication of chickenpox in children, very few children with chickenpox will develop flesh-eating disease.

Other Considerations

Chickenpox costs Canadians more than $122 million per year. This figure represents the cost of medical and hospital care, along with personal and productivity costs for parents and others who take time away from work to be caregivers.

Chickenpox Vaccine

The vaccine for chickenpox was licensed for use in Canada in 1998. It is given by needle, and is very safe. The side effects are temporary and usually mild. For example, some people have a sore spot or some tenderness where the needle went in. Up to 15% may have a mild fever that lasts for a few days. Up to 6% may develop a rash that resembles a mild case of chickenpox within a week or two of vaccination. The rash will clear up in about five days. Overall, these side effects are far less harmful than the potential complications from a serious case of chickenpox.

New Recommendations about Chickenpox Vaccine

The National Advisory Committee on Immunization (NACI) is a group of experts that provides Health Canada with ongoing and timely medical, scientific, and public health advice relating to immunization. In the *2002 Canadian Immunization Guide*, NACI recommends the chickenpox vaccine for healthy children (age 12 months and up), teenagers, and adults who have not already had chickenpox.

If you have had chickenpox once, you do not need to get the vaccine. But, a dose of the vaccine is unlikely to cause any harm as long as your overall general health is good.

However, NACI advises that certain people should *not* get the chickenpox vaccine, including:

- people who have a prior history of severe allergic reaction to the vaccine, or other components of the vaccine;
- people with weak immune systems, unless under the supervision of a specialist in infectious diseases;

- pregnant women or those who are trying to get pregnant; and
- babies less than a year old.

The cost of the chickenpox vaccine may or may not be covered by your health plan. Some provinces include it as part of their publicly funded immunization programs, while others have the matter under consideration.

Minimizing Your Risk

Obtain reliable information about chickenpox and chickenpox vaccine from credible sources. Talk to your doctor or health care provider about whether the chickenpox vaccine is right for you and your family.

Health Canada's Role

Health Canada regulates vaccines in Canada through a rigorous licensing process. This includes an extensive pre-market review of information about a vaccine's safety and effectiveness, and post-market assessment, such as tracking serious adverse reactions. In addition, Health Canada monitors and analyzes the incidence of vaccine-preventable diseases, develops guidelines for the control of diseases, and works with the provinces and territories on strategies to manage infectious diseases. Health Canada also participates in public awareness campaigns designed to help Canadians make informed decisions about immunization.

A.3 Sample Decision Guide from Mayo Clinic

Adjuvant therapy for breast cancer

Permission to reproduce this decision guide could not be obtained, but it is freely available online:
`http://www.mayoclinic.com/health/breast-cancer-treatment/AT99999`

Appendix B

Annotating Coreference in Health Education

This document describes the exact procedure and scheme used to annotate coreference in the health education corpus. Examples are given to help clarify the task, including ones directly from the corpus where possible.

Instructions for using the annotation tool are given first, followed by an overview of the general procedure. Specific guidelines about annotating markables and coreference are provided next. The document ends with a few reminders and a quick reference sheet.

Using PALinkA

Annotation is basically just adding special tags to existing documents. We use a special tool designed for this task called PALinkA [105]. PALinkA has a graphical user interface and shortcut keys to simplify the annotation task (e.g., by hiding unnecessary details).

The main tags used in our annotation scheme (and their associated shortcut keys) are as follows:

- `<MARKABLE>` (F9): This is the most common tag. Before coreference can be annotated, the NPs involved must be annotated as markables. Note that not all NPs are involved in coreference, so not all have to be annotated as markables. (However, it is fine to annotate more markables than strictly necessary.)

- <COREF> (F11): This tag indicates coreference, linking two markables together. It also has two special attributes (explained later) for recording additional information. A single markable cannot have more than one <COREF> tag.
- <UCOREF> (F12): This tag indicates probable coreference and is only used when there is some uncertainty (i.e., the text is ambiguous or the annotator is unsure). Adding an explanatory comment attribute is necessary for this kind of tag, whereas it is optional for all others.

Annotating a markable in PALinkA is done as follows:

1. Select the <MARKABLE> (F9) tag if not already selected[1].
2. Select the group of words to be annotated as a markable.
3. Press ENTER to confirm (or ESC to cancel).
4. Press ENTER again to skip adding a comment. (If there is something to be noted about this markable, do so before pressing ENTER.)

Adding a coreference annotation is also straightforward:

1. Select the <COREF> (or <UCOREF>) tag if not already selected (F11 or F12, respectively).
2. Click on a markable that corefers with another.
3. Click on the markable that corefers with the first (either directly in the text, or from the list on the right side of the interface).
4. Select a value for the relationship attribute (or just press ENTER to accept the default value of IDENT).
5. Select a value for the reference attribute (or just press ENTER to accept the default value of NP).
6. Press ENTER to skip adding a comment. (If there is something to be noted about this instance of coreference, do so before pressing ENTER.)

Finally, here's how to delete an annotation (of any type):

1. CTRL-click on a tag.

[1] The currently selected tag is shown on the bottom right of the interface.

2. If there is more than one tag at that position and a dialogue box appears, select which one you intend to delete.
3. Press DELETE to confirm or ESC to cancel.

Note that it is faster to overwrite an existing annotation with a new one if you plan on replacing it anyway.

General Procedure

The basic annotation procedure is as follows:

1. Open a new document in PALinkA. Note that noun phrases identified during preprocessing are already highlighted grey.
2. Briefly skim the document just to become familiar with its content, watching out for referring expressions and instances of coreference.
3. Annotate topic NPs (e.g., *asthma* in the "Asthma" article) all at once using the auto-tagging feature of PALinkA (found under EDIT > OPTIONS).
4. Carefully scan through the document word-by-word looking for coreference, using the suggested NPs as potential markables. When coreference is found:
   (a) Add markable annotations to the coreferring NPs.
   (b) Add a coreference link between the new markables.
5. Once at the end of the document, check it over to ensure nothing was missed. The list of markables on the right side of the interface is helpful in this way: look for 'forgotten' markables that are not in a chain but should be. Use the built-in search to quickly find possible antecedents.

If unclear about a specific case, either discuss it immediately[2] (if other annotators are available) with other annotator(s) or use the <UCOREF> tag and discuss it later. Obvious typos (e.g., *there* for *their*) should be treated as though the typo did not occur, with a comment to that effect inserted in the annotation.

PALinkA keeps track of how long each document is being annotated. When taking breaks or seeking clarification, remember to hit the 'Pause' button (found in the EDIT menu).

[2] Except during overlapping annotation for calculating inter-annotator agreement, of course.

Markables

The only referring expressions that we are interested in are noun phrases (NPs), and we only need to annotate those that actually participate in coreference. (Annotating extra markables is also fine, of course.)

In general, we annotate full noun phrases, i.e. the head of the NP along with all modifiers (prenominal and/or postnominal). The annotation may not 'skip' some or all modifiers of the head. So, for example, *respiratory disease* may not be annotated as *respiratory [disease]*.

However, in cases where a non-head NP is embedded within a larger NP, the embedded NP is markable if it occurs elsewhere in the text on its own (i.e., as the head of an NP). For example, the *chickenpox* within *chickenpox vaccine* is not itself markable unless *chickenpox* is the head of an NP somewhere else in the text. If so, the NP becomes *[[chickenpox] vaccine]* and then the two occurrences are linked by coreference.

The situation is the same for coordinated NPs: we do not separately annotate the involved NPs unless they occur separately elsewhere. So, for example, we would not mark *a doctor* in *a doctor or other licensed health care provider* unless it occurs somewhere else in the text on its own. Note that *his/her* and *his or her* (as in, e.g., *if the patient left his or her identification at home*) are usually best treated as a single markable.

For possessive forms, however, the embedded possessive is always markable (e.g., *[[your] condition]* and *[[Health Canada's] role]*).

Keep in mind that sometimes an embedded NP may refer to its enclosing NP, as with the *their* in *[homes that have [their] outside lights turned on]*.

In the case of named entities (e.g., organizations, persons, locations), substrings are not considered markable. So for *Health Canada*, the mention of *Canada* is not markable and therefore cannot corefer with other expressions. We also do not normally treat times and quantities as markable because of their rarity in health education.

Finally, note that we annotate everything in the documents, including titles, section headings and bulleted lists. Those parts that we do not annotate (e.g., tables of contents, 'More Info' sections) are removed during preprocessing.

Coreference

We annotate two levels of coreference: specific and generic. Specific coreference involves the exact same referent. For example, *these glands* and *the oily glands in* *his skin* in (1) both refer to glands belonging to someone named 'Arthur'. Generic coreference, on the other hand, describes cases like (2) where the referent is not a specific entity in the text, but rather a general concept or class of things.

(1) *Arthur's acne is caused by inflammation of* ***the oily glands in his skin****.* *When the ducts of* ***these glands*** *become blocked by layers of skin . . .*

(2) *Acne is caused by inflammation of* ***the oily glands in the skin****. When the ducts of* ***the glands*** *become blocked by layers of skin . . .*

When adding a new NP to a coreference chain, it usually does not matter which element already in the chain is chosen as the antecedent, but the direction always matters: anaphors should only be annotated as pointing backward in the text, except in the (rare) situation of cataphora. For marking coreference relations other than identity, the antecedent chosen does matter. Consider the following series of referring expressions:

(A) *a doctor*
(B) *the physician*
(C) *the doctor*

(B) is a synonym of (A). But (C) could either be annotated as a synonym of (B) or identical to (A). In such cases, identity is to be preferred. In other words, if there is an identical antecedent available, always choose it over a synonym, generalization, etc.

Note also that coreference should be textually certain in order to be annotated. For example, the two NPs in (3) may or may not corefer, so no coreference should be annotated.

(3) *Hippocrates* ***may*** *be the greatest physician of all time.*

Similarly, there are cases where coreference is based on a person or organization's (possibly biased) point-of-view (e.g., *Under CEPA, biotechnology is defined as "the application of . . . "*). Since this is only possible coreference, it should not be annotated. Likewise, we do not annotate what was only true in the past (e.g., *She was a doctor-in-training*).

The annotation of demonstrative, relative and indefinite anaphors such as (4)–(6) is optional as they are not a high priority for this iteration of the project.

(4) *The controls could also include* ***an outright ban****, but* ***this*** *is rarely used.*

(5) *Cribs should have* ***a firmly fixed mattress support****, the assembly of* ***which*** *usually requires tools.*

(6) *. . . may indicate* ***a health condition*** *when* ***none*** *is present.*

Types of relations

While other types of coreference certainly exist, we only annotate the most well-studied varieties in order to make the annotation task feasible.

Identity

We mainly annotate markables as coreferential when their referents are identical. For this reason, IDENT is the default relation that comes up when adding a coreference annotation in PALinkA. This is the proper relation for cases where the head noun is present in both NPs (e.g., *the medical treatment . . . treatments*), for pronouns (e.g., *tumours . . . they*) and for proper nouns (e.g., *Health Canada . . . it*).

The identity relation is applied more strictly for generic coreference: the heads of the referring expressions usually need to be identical, not just overlapping. So, for example, there is no identity relation between generic references to *red marks* and *marks* (indeed, this is a different kind of coreference altogether, which we do not consider) but there would be for specific reference like *red marks . . . the marks.*

The identity relation is symmetric, i.e., for two referents A and B, if A is identical to B then B is also identical to A. It is also transitive, so if A is identical to B and B is identical to C, then A is identical to C. This means that every expression within a coreference chain corefers with every other in the chain. This is unlike the part/whole relation, for example.

Note that IDENT is the appropriate relation to use for acronyms (e.g., *STI* for *sexually transmitted infection*), not SYNONYM.

Synonymy

If two NPs that are synonyms corefer, the SYNONYM relation should be selected in PALinkA when adding coreference. Note that the heads may be different but the meanings (and levels of detail) must essentially be the same, as in the following examples:

(7) *doctor . . . physician*

(8) *women of childbearing age . . . female patients of childbearing age*

Other

The remaining relations of interest are generalization and specialization, which are cases where anaphors are less or more specific (respectively) than their antecedents. Both of these relations are exemplified below. Note that both are just marked OTHER for simplicity, but a comment should be included that specifies which one applies.

(9) Generalization: *asthma . . . the disease*

(10) Specialization: *three cities . . . Cambridge, Kitchener and Waterloo*

Unlike IDENT and SYNONYM, this kind of relation only applies to specific coreference. So, for example, generic referring expressions like *asthma* and *disease* (note the lack of definite article *the*) are not considered to be coreferential with one another: *asthma* refers to a singular disease, whereas *disease* refers to the concept of disease in general (this would be indirect coreference, which we do not annotate).

To further illustrate the difference, consider *oral antibiotics* and *antibiotics*. While the second might seem to be a generalization of the first, in fact the second is referring generically to a class of things (antibiotics) and the first is generically referring to a subset (oral antibiotics). For this reason, these would also not be marked as coreferential. However, if the example were changed to be specific (e.g., *oral antibiotics . . . these antibiotics*) we would then consider it to be coreference.

Sometimes the coreference is clear but the type is not, and choosing between IDENT, SYNONYM and OTHER is difficult. For example, consider the following variety of referring expressions, all of which (in context) refer to the same thing:

medical test kits for home use
medical test kits for home
medical kits for self-testing
medical test kits
home-use test kits
home-use medical tests
do-it-yourself medical tests
do-it-yourself tests
test kits for home use
test kits
tests
kits

A useful heuristic is to look at the head of the NP. If an anaphor and antecedent have overlapping heads, treat it as IDENT regardless of the level of detail. In the previous example, all variations could therefore be considered IDENT because the head is *test kits*. If the heads are not equal, but the level of detail is the same, consider it SYNONYM. Finally, if the heads are not equal and the levels of detail differ, consider it OTHER (generalization/specialization). For this example, a markable generalization would be, e.g., *test kit for home use ... the product*.

As another example, notice that *the Prime Minister ... Stephen Harper* would be specialization (the heads being different, with the second more specific than the first) whereas *Harper ... PM Stephen Harper* would not be because of overlapping heads.

It has been mentioned that we are not interested in annotating so-called indirect coreference. Here are some additional examples of what **not** to annotate:

- Set/subset

  (11) *Asthma appears to result from the interaction of* ***a number of factors****, including:* ***predisposing factors****,* ***causal factors****, and* ***contributing factors****.*
  The three categories of factors generically refer to subsets of *a number of factors*, and therefore don't corefer.

  (12) **Car seats** must carry compliance labels... **Car seats that are cracked or broken**.
  Damaged car seats are a subset of all car seats.

- Class/instance

  (13) *...* ***antibiotics*** *(such as* ***tetracycline*** *and* ***erythromycin****)*
  Tetracycline and erythromycin are instances of antibiotics.

  (14) ...**causal factors**, which may sensitize the airways (e.g. **dust mites**, **cockroaches** or **workplace contaminants**)
  Similar to above.

- Part/whole

  (15) ***The appendix*** *is completely unnecessary for the overall health of* ***the body****.*
  An appendix is part of the body.

Types of reference

There are five different types of reference that we annotate. The vast majority of cases are NP.

NP

This type of reference is the default when annotating a new instance of coreference, and should be selected when none of the others apply. The label NP is a bit misleading because all coreference that we annotate involves NPs.

We do not annotate coreference involving verbs, adverbs or clauses. For example, in (16) the *it* refers to a clause (equivalent to *they have whooping cough*). Because the coreference is not between NPs, the *it* is not annotated as coreferential for our purposes, however it is still considered markable. In this case, a comment should be added explaining the situation.

(16) *Older members of a household may have whooping cough without even realizing **it**.*

Some more examples of non-NP coreference that we do **not** annotate:

(17) *See your doctor if you are feeling sick or worried, or if the test instructions recommend **it**.*

(18) *Because tiny particles from volcanic ash can go deep into your lungs, you should not breathe it if you can help **it**.*

(19) *During an attack, the muscles around the airways can tighten and the airways can produce mucus. **These conditions** make it even harder to breathe.*

Appositives

Appositive phrases (usually set off by commas) should be marked as coreferential and specified as APPOSITIVE (instead of NP). For example, in (20) *Bacne* and the appositive phrase *acne occurring specifically on the back* should be marked as coreferential. Because the appositive phrase uniquely identifies *Bacne*, the relation type is IDENT.

(20) *Bacne, **acne occurring specifically on the back**, ...*

On the other hand, the appositive phrase in (21) (*a common skin condition*) could equally apply to sunburn or eczema. Because it is more general, the relation type is OTHER (generalization).

(21) *Acne, **a common skin condition**, can be treated in many ways.*

Finally, negative appositions (e.g., *Acne, never a self-esteem booster, ...*) should not be annotated as coreferential.

Copula (linking verbs)

When the subject NP of a sentence is equated (or associated) with a predicate NP using a copula (sometimes called a linking verb), the NPs are tagged as being coreferential (reference type COPULAR). For example, in (22) the NPs *Nova Scotia* and *the province with the highest cancer rates in the country* are coreferential because of the verb *is.* Other copula include *become, seem, look, appear, remain*, and *stay.* Basically, if the verb clearly implies equivalence, it can be considered copula.

(22) *Nova Scotia* ***is*** *the province with the highest cancer rates in the country.*

(23) *They* ***are called*** *sebaceous glands*

Expressions of equivalence using *as* (e.g., *Tylenol* ***is known as*** *the most popular pain relief drug*) or *i.e.* (e.g., *asthma symptoms and attacks,* ***i.e.*** *episodes of more severe shortness of breath*) are also marked as copula.

Note that copular coreference is considered to be the IDENT relation when the predicate uniquely identifies the subject (as in (24)) but OTHER (generalization or specialization) when the predicate is more or less general than the subject (25).

(24) *Chlamydia is the most common bacterial STI in Canada.*

(25) *Asthma is an important factor in school absences in children.*

As with appositives, negative copula (e.g., *Tylenol is* ***not*** *. . .*) should not be annotated as coreferential.

Brackets and dashes

NPs in brackets or between dashes following an NP can also be coreferential, and should be specified as BRACKETED and DASHED, respectively. Consider the following examples:

(26) *Trigeminal neuralgia* ***(TN)*** *is a painful disorder of the nervous system.*

(27) *What many Canadians do not realize is that cancer—****the leading cause of premature death****—is mostly preventable through healthy living.*

(28) *Asthma symptoms and attacks (i.e.* ***episodes of more severe shortness of breath)*** *. . .*

These types of reference generally involve the IDENT relation.

Cataphora

A cataphoric reference is when an NP corefers with an expression introduced for the first time later in the text (i.e., it refers *forward*). This should be marked as coreference with the reference type CATAPHORIC. For example:

(29) *Although **she** never said so, **the woman** was already three months pregnant.*

Note that this phenomenon is quite rare in health articles (and, indeed, most corpora).

Things to keep in mind

- Some pronouns, called 'dummy' or 'pleonastic', do not actually refer to anything. For example, the *it* in *it is important to brush your teeth regularly* really does not refer to anything, and therefore cannot participate in coreference. The only dummy pronouns likely to be encountered are occurrences of *it*. While annotating, be certain to mark such pronouns anyway in order to leave a comment.

- Second-person pronouns (e.g., *you* and *your*) should be marked as coreferential. One way to do this is to link all instances of *you* with one another, and all instances of *your* with one another, and then finally link one instance of *you* with *your* (or vice versa) to connect the two chains.

  First-person pronouns (especially *we* and *our*) are also usually coreferential. In some cases, their antecedents are nearby in the text (e.g., ***Canadians** generally support the biotechnology revolution, but there are concerns that it might affect **our** health*).

- Sometimes an NP has multiple referents as in *wait until both you and your partner have completed your treatment* where the second *your* refers to both *you* and *your partner*. In such cases, PALinkA allows them to both be marked with ADD REFERENT (from the TOOLS menu).

- Watch out for coordinated NPs. When an NP such as *the drug's cost or side effects* is followed by a reference to *the drug's side effects* — a tricky situation because *side effects* is separated from *the drug's* — we just mark the second mention as coreferential with the first mention of *side effects* and include a comment mentioning a coordinated NP.

- Watch out for shifts between specific and generic reference as in the following:

(30) ***An organism** is a body of living matter. . . . The kind of information required to evaluate potential hazards includes: the identity of **the organism** . . .*

The context makes it clear that the second referring expression refers to a specific organism (i.e., one involved with a new biotechnology product), whereas the first refers to organisms in general. Therefore the two do not corefer.

Sometimes there can be generic coreference even when the NPs suggest different levels of detail:

(31) *Make sure **indoor and outdoor decorative lights** are certified. . . . Check **lights** for broken or cracked sockets, frayed or bare wires or loose connections.*

In this case it is clear from the context that the second occurrence of *lights* is also referring at same level of detail (e.g., not flash lights), so there is still generic coreference. This only seems to happen with sentences in close proximity, and only with less detailed NPs referring to more detailed ones.

- Omitted words are never marked as coreferential, i.e., we do not mark 'zero anaphora'. For example, if the words *of acne* are omitted in the sentence *In severe cases [of acne], treatment may be needed*, the ellided *acne* is not marked as coreferential with other references to acne.

- Perspective shifts are also problematic. For example, It's Your Health articles often begin in the third person (e.g., *people with asthma . . . their symptoms*) and then switch to second person (e.g., *your symptoms*), especially in "Minimizing Your Risk" sections. This shift 'interrupts' coreference because *their* symptoms are not necessarily *your* symptoms.

- NPs in bulleted lists often corefer with NPs introducing the list, but this can be tricky to notice because lists are not properly displayed in PALinkA. For example:

  (32) *When buying **a test**: . . . Remember, buying **an unlicensed test** increases the chances that **the test** is unreliable. Read and follow the directions for storing **the test**.*

  In this case, the first instance of *the test* corefers with *an unlicensed test*, whereas the second instead corefers with *a test*. The last two sentences are actually separate bullet points, despite the fact that they appear to be consecutive sentences. In such cases, it can be helpful to check the original document.

  Similarly, NPs in section headings often become antecedents for anaphors within such sections.

- The part-of-speech tagger used for preprocessing is not perfect. Sometimes words are identified as nouns that are not, and occasionally nouns are missed. Use your best judgment.
- Besides occasional typos, there are other editing problems like repeated sentences (especially in EatRight articles). When a sentence or section is duplicated, only annotate the first occurrence.

Quick Reference

For each noun *N*,

1. Identify containing noun phrase *NP* (with pre-determiners, determiner, post-determiners, head noun, post-nominals, conjunctions), if any
   - If this NP has already been considered, skip to next *N*
2. Determine whether *NP* corefers with any other markables
   - Check for specific coreference. (Recall that this is normally only the case for definite NPs, and that pronouns are normally not antecedents.)
   - Else. . .
     - Check if coreference type is apposition, copula, brackets or dashes
     - Check for coreference with an NP in the title or section heading
     - Check for generic coreference
   - If no coreference, but *N* is not the head, treat *N* alone as *NP* and check again
   - If still no coreference, check if typo or dummy pronoun
     - If typo, treat as if typo corrected and add comment
     - If dummy, add comment and continue but skip step 4
     - Else skip to next *N*
3. Add markable annotation to *NP*
   - If first mention (e.g., indefinite), skip to next *N*
4. Add coreference annotation between *NP* and antecedent
   - Select *NP*
   - Select antecedent *A*
     - If *NP* is IDENT to any available antecedent, select that antecedent
     - Else (SYNONYM or OTHER) just select first mention
   - Select relation of *NP* to *A*
     - IDENT if *NP* and *A* are identical
     - SYNONYM if *NP* has different head than *A* but similar meaning
     - OTHER if *NP* is more general or specific than *A* (add comment specifying relation)
   - Select reference type of *NP*
     - APPOSITIVE if *NP* is set off by commas
     - COPULAR if *NP* follows verb suggesting equivalence (e.g., *is, are*)
     - BRACKETED if *NP* is within brackets
     - DASHED if *NP* is set off with dashes
     - CATAPHORIC if *A* is later in text than *NP*
     - NP otherwise

References

[1] S. Adolphs, B. Brown, R. Carter, C. Crawford, and O. Sahota. Applying corpus linguistics in a health care context. *Journal of Applied Linguistics*, 1(1):9–28, 2004.

[2] C. Aone and S. Bennett. Discourse tagging tool and discourse tagged multilingual corpora. In *Proceedings of the International Workshop on Sharable Natural Language Resources (SNLR)*, pages 71–77, Ikoma, Japan, August 1994.

[3] C. Aone and S. Bennett. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 122–129, Cambridge, Massachusetts, June 1995.

[4] D.E. Appelt, J.R. Hobbs, J. Bear, D. Israel, M. Kameyama, A. Kehler, D. Martin, K. Myers, and M. Tyson. SRI International FASTUS system: MUC-6 test results and analysis. In *Proceedings of the 6th Message Understanding Conference (MUC)*, pages 237–248, Columbia, Maryland, November 1995.

[5] R. Artstein and M. Poesio. Kappa^3 = alpha (or beta). Technical Report CSM-437, University of Essex Department of Computer Science, 2005.

[6] B. Baldwin. CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, Madrid, Spain, July 1997.

[7] B. Baldwin, J. Reynar, M. Collins, J. Eisner, A. Ratnaparki, J. Rosenzweig, A. Sarkar, and S. Bangalore. Description of the University of Pennsylvania system used for MUC-6. In *Proceedings of the 6th Message Understanding Conference (MUC)*, pages 177–191, Columbia, Maryland, November 1995.

[8] C. Barbu. Automatic learning and resolution of anaphora. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 22–27, Tzigov Chark, Bulgaria, September 2001.

[9] C. Barbu. *Bilingual Pronoun Resolution: Experiments in English and French.* PhD thesis, University of Wolverhampton, 2003.

[10] C. Barbu and R. Mitkov. Evaluation tool for rule-based anaphora resolution methods. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 34–41, Toulouse, France, July 2001.

[11] D. Bean and E. Riloff. Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 373–380, College Park, Maryland, June 1999.

[12] D. Biber. *Variation across speech and writing.* University Press, Cambridge, UK, 1988.

[13] D. Biber. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, 5:257–269, 1990.

[14] D. Biber and E. Finegan. Intra-textual variation within medical research articles. In *Corpus-based research into language*, pages 201–222. Rodopi, Amsterdam, 1994.

[15] Reference guide for the British National Corpus. `http://www.natcorp.ox.ac.uk/docs/userManual`, October 2000.

[16] D. Byron. The uncommon denominator: A proposal for consistent reporting of pronoun resolution results. *Computational Linguistics, Special Issue on Computational Anaphora Resolution*, 27(4):569–577, 2001.

[17] Health Canada. It's Your Health. `http://www.healthcanada.gc.ca/iyh`, November 2007.

[18] J. Carbonell and R. Brown. Anaphora resolution: A multi-strategy approach. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING)*, pages 96–101, Budapest, Hungary, August 1988.

[19] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[20] E. Charniak. Toward a model of children's story comprehension. Technical Report AI TR-266, Massachusetts Institute of Technology Artificial Intelligence Laboratory, October 2006.

[21] N. Chinchor. Overview of MUC-7. In *Proceedings of the 7th Message Understanding Conference (MUC)*, Fairfax, Virginia, April 1998.

[22] N. Chinchor. 7th Message Understanding Conference (MUC). LDC2001T02, Linguistic Data Consortium, Philadelphia, 2001.

[23] N. Chinchor and B. Sundheim. 6th Message Understanding Conference (MUC). LDC2003T13, Linguistic Data Consortium, Philadelphia, 2003.

[24] Mayo Clinic. Treatment Decisions. `http://www.mayoclinic.com/health/TreatmentDecsionIndex/TreatmentDecisionIndex`, November 2007.

[25] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[26] D. Connolly, J. Burger, and D. Day. A machine learning approach to anaphoric reference. In *Proceedings of the International Conference on New Methods in Language Processing (NEMLAP)*, pages 255–261, Manchester, UK, September 1994.

[27] R. Crawley, R. Stevenson, and D. Kleinman. The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, 4:245–264, 1990.

[28] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)*, pages 168–175, Philadelphia, Pennsylvania, 2002.

[29] M. Davis, T. Dunning, and B. Ogden. Text alignment in the real world: Improving alignments of noisy translations using common lexical features, string matching strategies and n-gram comparisons. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 67–74, Dublin, Ireland, March 1995.

[30] T.C. Davis, E.J. Mayeaux, D. Fredrickson, J.A. Bocchini, Jr., R.H. Jackson, and P.W. Murphy. Reading ability of parents compared with reading level of pediatric patient education materials. *Journal of Pediatrics*, 93:460–468, 1994.

[31] D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, and M. Vilain. Mixed-initiative development of language processing systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 348–355, Washington, DC, April 1997.

[32] L. Deléger, M. Merkel, and P. Zweigenbaum. Enriching medical terminologies: an approach based on aligned corpora. In *Proceedings of Medical Informatics Europe*, pages 747–752, Maastricht, Netherlands, August 2006.

[33] L. Deléger, M. Merkel, and P. Zweigenbaum. Using word alignment to extend multilingual medical terminologies. In *Proceedings of the LREC Workshop on Acquiring and Representing Multilingual, Specialized Lexicons: The Case of Biomedicine*, pages 9–14, Genoa, Italy, May 2006.

[34] B. Di Eugenio and M. Glass. The Kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101, 2004.

[35] C. DiMarco, H.D. Covvey, P. Bray, D.D. Cowan, V. DiCiccio, E. Hovy, J. Lipa, and D. Mulholland. The development of a natural language generation system for personalized e-health information (poster presentation). In *Proceedings of the 12th International Health (Medical) Informatics Congress (Medinfo)*, pages 2339–2340, Brisbane, Australia, August 2007.

[36] C. DiMarco, D.D. Cowan, P. Bray, H.D. Covvey, V. DiCiccio, E. Hovy, J. Lipa, and D. Mulholland. A physician's authoring tool for generation of personalized health education in reconstructive surgery. In *Proceedings of the American Association for Artificial Intelligence (AAAI) Spring Symposium on Argumentation for Consumers of Healthcare*, pages 39–46, Stanford, California, March 2006.

[37] M. Dimitrov, K. Bontcheva, H. Cunningham, and D. Maynard. A light-weight approach to coreference resolution for named entities in text. In A. Branco, T. McEnery, and R. Mitkov, editors, *Anaphora Processing: Linguistic, Cognitive and Computational Modelling.* John Benjamins Publishing, Amsterdam, 2004.

[38] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. The Automatic Content Extraction (ACE) program: Tasks, data, and evaluation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 837–840, Lisbon, Portugal, May 2004.

[39] B. Di Eugenio. On the usage of Kappa to evaluate agreement on coding tasks. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 441–444, Athens, Greece, May 2000.

[40] R. Evans. A comparison of rule-based and machine learning methods for identifying non-nominal *It*. In *Proceedings of the Second International Conference on Natural Language Processing*, pages 233–241, Patras, Greece, June 2000.

[41] R. Evans. Applying machine learning toward an automatic classification of *It*. *Literary and Linguistic Computing*, 16(1):45–57, 2001.

[42] G. Eysenbach and C. Köhler. Health-related searches on the Internet. *Journal of the American Medical Association*, 291(24):2946, 2004.

[43] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Massachusetts, 1998.

[44] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233, 1948.

[45] R. Flesch. *The art of readable writing.* Harper, New York, 1949.

[46] R. Flesch. *How to write plain English.* Harper and Row, New York, 1979.

[47] S. Fox. Online health search 2006. Pew Internet & American Life Project (`http://www.pewinternet.org/pdfs/PIP_Online_Health_2006.pdf`), October 2006.

[48] W.N. Francis. A standard sample of present-day English for use with digital computers. Report to the U.S Office of Education on Cooperative Research Project No. E-007, 1964.

[49] M.C. Freda. The readability of American Academy of Pediatrics patient education brochures. *Journal of Pediatric Health Care*, 19(3):151–156, 2005.

[50] J. Frederiksen. Understanding anaphora: Rules used by readers in assigning pronominal referents. *Discourse Processes*, 4:323–347, 1981.

[51] D.B. Friedman, L. Hoffman-Goetz, and J.F. Arocha. Readability of cancer information on the internet. *Journal of Cancer Education*, 19:117–122, 2004.

[52] S.C. Garrod, D. Freudenthal, and E. Boyle. The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language*, 33:39–680, 1994.

[53] S.C. Garrod and A.J. Sanford. Interpreting anaphoric relations: The integration of semantic information while reading. *Journal of Verbal Learning and Verbal Behavior*, 16:77–90, 1977.

[54] R. Garside, S. Fligelstone, and S. Botley. Discourse annotation: anaphoric relations in corpora. In R. Garside, G. Leech, and A. McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 66–84. Longman, London, 1997.

[55] C. Gasperin, S. Salmon-Alt, and R. Vieira. How useful are similarity word lists for indirect anaphora resolution? In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC)*, Sao Miguel, Azores, September 2004.

[56] N. Ge. Annotating the Penn Treebank with coreference information. Technical report, Brown University, Department of Computer Science, 1998.

[57] M.A. Gernsbacher and D.J. Hargreaves. Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, 27:699–717, 1988.

[58] B.J. Grosz and C.L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[59] J.K. Gundel, N. Hedberg, and R. Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307, 1993.

[60] S.M. Harabagiu, R.C. Bunescu, and S. Trausan-Matu. COREFDRAW - a tool for annotation and visualization of coreference data. In *Proceedings of the 13th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 273–279, Dallas, Texas, November 2001.

[61] T.A. Harley. *Psychology of Language*. Psychology Press, New York, second edition, 2001.

[62] L. Hasler, C. Orăsan, and K. Naumann. NPs for events: Experiments in coreference annotation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1167–1172, Genoa, Italy, May 2006.

[63] E. Hinrichs, S. Kübler, and K. Naumann. A unified representation for morphological, syntactic, semantic, and referential annotations. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 13–20, Ann Arbor, Michigan, June 2004.

[64] L. Hirschman, P. Robinson, J. Burger, and M. Vilain. Automating coreference: The role of annotated training data. In *Proceedings of the American Association for Artificial Intelligence (AAAI) Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 1419–1422, Stanford, California, 1997.

[65] G. Hirst. *Anaphora in Natural Language Understanding: A Survey*, volume 119 of *Lecture Notes in Computer Science*. Springer, Berlin, 1981.

[66] J.R. Hobbs. Resolving pronoun references. *Lingua*, 44:311–338, 1978.

[67] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the 7th Message Understanding Conference (MUC)*, Fairfax, Virginia, April 1998.

[68] J. Järvikivi, R.P.G. van Gompel, J. Hyönä, and R. Bertram. Ambiguous pronoun resolution: Contrasting the first-mention and subject preference accounts. *Psychological Science*, 16:260–264, 2005.

[69] D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ, 2000.

[70] M. Kameyama. Recognizing referential links: An information extraction perspective. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53, Madrid, Spain, July 1997.

[71] C. Kennedy and B. Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 113–118, Copenhagen, Denmark, August 1996.

[72] R.N. Khurana, P.P. Lee, and P. Challa. Readability of ocular medication inserts. *Journal of Glaucoma*, 12:50–53, 2003.

[73] R. Kibble and K. van Deemter. Coreference annotation: Whither? In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1281–1286, Athens, Greece, May 2000.

[74] J.P. Kincaid, R.P. Fishburne, Jr., R.L. Rogers, and B.S. Chissom. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Research Branch Report 8-75, Millington, TN: Naval Technical Training, U.S. Naval Air Station, Memphis, TN, 1975.

[75] O. Kirksey, K. Harper, S. Thompson, and M. Pringle. Assessment of selected patient educational materials of various chain pharmacies. *Journal of Health Communication*, 9:91–93, 2004.

[76] B. Klebanov. Using latent semantic analysis for pronominal anaphora resolution. Master's thesis, University of Edinburgh, 2001.

[77] O. Krasavina and C. Chiarcos. PoCoS - Potsdam coreference scheme. In *Proceedings of the ACL Linguistic Annotation Workshop*, pages 156–163, Prague, Czech Republic, June 2007.

[78] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology.* Sage, Newbury Park, CA, 1980.

[79] S. Lappin and H. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

[80] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, Massachusetts, 1999.

[81] D. Marcu, E. Amorrortu, and M. Romera. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*, pages 48–57, College Park, Maryland, June 1999.

[82] M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[83] A. McCallum and B. Wellner. Toward conditional models of identity uncertainty with application to proper noun coreference. In *Proceedings of the IJCAI Workshop on Information Integration on the Web*, pages 79–84, Acapulco, Mexico, August 2003.

[84] A.M. McEnery and Z. Xiao. Domains, text types, aspect marking and English-Chinese translation. *Languages in Contrast*, 2(2):211–229, 2002.

[85] D. McKelvie, A. Isard, A. Mengel, M.B. Moller, M. Gross, and M. Klein. The MATE workbench — an annotation tool for XML coded speech corpora. *Speech Communication*, 33(1-2):97–112, 2001.

[86] G.H. McLaughlin. SMOG grading: A new readability formula. *Journal of Reading*, 12(8):639–646, 1969.

[87] R.C. Miller and K. Bharat. SPHINX: A framework for creating personal, site-specific web crawlers. In *Proceedings of the 7th World Wide Web Conference*, pages 119–130, Brisbane, Australia, April 1998.

[88] R. Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 867–875, Montreal, Canada, August 1998.

[89] R. Mitkov. Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC)*, pages 96–107, Lancaster, UK, November 2000.

[90] R. Mitkov. Outstanding issues in anaphora resolution (invited talk). In *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 110–125, Berlin, Germany, February 2001.

[91] R. Mitkov. *Anaphora Resolution.* Longman, London, 2002.

[92] R. Mitkov. Anaphora resolution. In R. Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, pages 266–283. University Press, 2003.

[93] R. Mitkov, B. Boguraev, and S. Lappin. Introduction to the Special Issue on Computational Anaphora Resolution. *Computational Linguistics*, 27(4):473–477, 2001.

[94] R. Mitkov, R. Evans, and C. Orăsan. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 168–186, Mexico City, February 2002.

[95] R. Mitkov, R. Evans, C. Orăsan, C. Barbu, L. Jones, and V. Sotirova. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC)*, pages 49–58, Lancaster, UK, November 2000.

[96] R. Mitkov and C. Hallett. Comparing pronoun resolution algorithms. *Computational Intelligence*, 23(2):262–297, 2007.

[97] R. Mitkov, C. Orăsan, and R. Evans. The importance of annotated corpora for NLP: the cases of anaphora resolution and clause splitting. In *Proceedings of the TALN Workshop "Corpora and NLP: Reflecting on Methodology Workshop"*, pages 60–69, Cargèse, Corsica, July 1999.

[98] T. Morton and J. LaCivita. WordFreak: An open tool for linguistic annotation (demo). In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL)*, pages 17–18, Edmonton, Canada, May 2003.

[99] MUC-6. Coreference Task Definition (version 2.3). In *Proceedings of the 6th Message Understanding Conference (MUC)*, Columbia, Maryland, November 1995.

[100] MUC-7. Coreference Task Definition (version 3.0). In *Proceedings of the 7th Message Understanding Conference (MUC)*, Fairfax, Virginia, April 1997.

[101] C. Müller and M. Strube. Annotating anaphoric and bridging relations with MMAX. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 90–95, Aalborg, Denmark, September 2001.

[102] G.L. Murphy. Processes of understanding anaphora. *Journal of Memory and Language*, 24:290–303, 1985.

[103] Y. Nishina. A corpus-driven approach to genre analysis: The reinvestigation of academic, newspaper and literary texts. *Empirical Language Research*, 1, 2007.

[104] Ministry of Health Promotion (Ontario). EatRight Ontario. `http://www.eatrightontario.ca`, November 2007.

[105] C. Orăsan. PALinkA: A highly customizable tool for discourse annotation. In *Proceedings of the Fourth SIGdial Workshop on Discourse and Dialogue*, pages 39–43, Sapporo, Japan, July 2003.

[106] C. Orăsan, D. Cristea, R. Mitkov, and A. Branco. Anaphora Resolution Exercise: An overview. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco, May 2008.

[107] C. Orăsan, R. Evans, and R. Mitkov. Enhancing preference-based anaphora resolution with genetic algorithms. In *Proceedings of the Second International Conference on Natural Language Processing*, pages 185–195, Patras, Greece, June 2000.

[108] C.D. Paice and G.D. Husk. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun ‘it’. *Computer Speech and Language*, 2:109–132, 1987.

[109] M. Palomar, A. Ferrandez, L. Moreno, P. Martinez-Barco, J. Peral, M. Saiz-Noeda, and R. Mufioz. An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*, 27(4):545–567, 2001.

[110] K. Parsons. An authoring tool for customizable documents. Master’s thesis, University of Waterloo, 1997.

[111] R. Passonneau. Applying reliability metrics to co-reference annotation. Technical Report CUCS-025-03, Columbia University, 1997.

[112] R. Passonneau. Computing reliability for coreference annotation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1503–1506, Lisbon, Portugal, May 2004.

[113] J.P. Pestian, L. Itert, and W. Duch. Development of a pediatric text-corpus for part-of-speech tagging. In *Proceedings of the Intelligent Information Processing and Web Mining Conference (IIPWM)*, pages 219–226, Zakopane, Poland, May 2004.

[114] M. Poesio and R. Artstein. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83, Ann Arbor, Michigan, June 2005.

[115] M. Poesio, F. Bruneseaux, and L. Romary. The MATE meta-scheme for coreference in dialogues in multiple languages. In *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*, pages 65–74, College Park, Maryland, June 1999.

[116] S.P. Ponzetto and M. Strube. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL)*, pages 192–199, New York City, June 2006.

[117] A. Popescu-Belis, L. Rigouste, S. Salmon-Alt, and L. Romary. Online evaluation of coreference resolution. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1507–1510, Lisbon, Portugal, May 2004.

[118] J. Preiss, C. Gasperin, and T. Briscoe. Can anaphoric definite descriptions be replaced by pronouns? In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1499–1502, Lisbon, Portugal, May 2004.

[119] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In *Proceedings of the 3rd ACL Workshop on Very Large Corpora*, pages 82–94, Cambridge, Massachusetts, June 1995.

[120] S.H. Rankin and K.D. Stallings. *Patient education: Principles and practices.* Lippincott, Philadelphia, fourth edition, 2001.

[121] B.K. Redman. *The practice of patient education.* Mosby, St. Louis, eighth edition, 1997.

[122] R. Reppen, N. Ide, and K. Suderman. American National Corpus (ANC) Second Release. LDC2005T35, Linguistic Data Consortium, Philadelphia, 2005.

[123] B. Russell. On denoting. *Mind*, 14:479–493, 1905.

[124] H. Schmid. Probabilistic part-of-speech tagging using decision trees. Technical report, IMS, University of Stuttgart, 1994.

[125] M. Scott. WordSmith Tools Version 5. Lexical Analysis Software, Liverpool, 2008.

[126] A. Sheldon. The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior*, 13:272–281, 1974.

[127] S. Siegel and N.J. Castellan, Jr. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, New York, second edition, 1988.

[128] J. Skelton and F.D.R. Hobbs. Concordancing: The use of language-based research in medical communication. *The Lancet*, 353:108–111, 1999.

[129] J. Skelton, A.M. Wearn, and F.D.R. Hobbs. 'I' and 'we': A concordancing analysis of how doctors and patients use first person pronouns in primary care consultations. *Journal of Family Practice*, 19(5):484–488, 2002.

[130] W. Soon, H. Ng, and D. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[131] C.M. Sperberg-McQueen and L. Burnard, editors. *TEI P4: Guidelines for Electronic Text Encoding and Interchange.* Text Encoding Initiative Consortium, Oxford, 2002.

[132] M. Strube, S. Rapp, and C. Müller. The influence of minimum edit distance on reference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 312–319, Philadelphia, Pennsylvania, July 2002.

[133] R. Stuckardt. Machine-learning-based vs. manually designed approaches to anaphor resolution: The best of two worlds. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC)*, pages 211–216, Lisbon, Portgual, September 2002.

[134] B. Sundheim. Overview of results of the MUC-6 evaluation. In *Proceedings of the 6th Message Understanding Conference (MUC)*, pages 13–31, Columbia, Maryland, November 1995.

[135] M. Templin. *Certain Language Skills in Children*. Greenwood, Westport, CT, 1957.

[136] J. Tetreault. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520, 2001.

[137] J. Thomas and A. Wilson. Methodologies for studying a corpus of doctor-patient interaction. In J. Thomas and M. Short, editors, *Using Corpora for Language Research*. Longman, London, 1996.

[138] O. Uryupina. *Knowledge Acquisition for Coreference Resolution*. PhD thesis, Saarland University, 2007.

[139] K. van Deemter and R. Kibble. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637, 2000.

[140] R. Vieira and M. Poesio. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593, 2000.

[141] R. Vieira and M. Poesio. Processing definite descriptions in corpora. In S. Botley and M. McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*, pages 189–212. John Benjamins Publishing, Amsterdam, 2000.

[142] R. Viera. *Definite Description Processing in Unrestricted Text*. PhD thesis, University of Edinburgh, 1998.

[143] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC)*, pages 45–52, Columbia, Maryland, November 1995.

[144] L.S. Wallace and E.S. Lennon. American Academy of Family Physicians patient education materials: Can patients read them? *Family Medicine*, 36:571–574, 2004.

[145] A.B. Zion and J. Aiman. Level of reading difficulty in the American College of Obstetricians and Gynecologists patient education pamphlets. *Obstetrics and Gynecology*, 74:955–960, 1989.

[146] G.K. Zipf. *Human behavior and the principle of least effort.* Addison-Wesley, Cambridge, Massachusetts, 1949.