Måns Huldén

Linguistic Complexity in Two Major American Newspapers and The Associated Press Newswire, 1900–2000

Pro-gradu avhandling i engelska
språket och litteraturen
Handledare: Tuija Virtanen-Ulfhielm
Åbo Akademi
Åbo 2004

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> *The Admiralty steadfastly professes its inability to throw any light on the situation, and there is reason to believe that the profession is made in good faith, at least by all but the very highest officers.*
> *—AP, Apr. 11, 1905*
>
> *He refused to comment further.*
> *— AP, Nov. 9, 1996*

In Jules Verne's *Around the World in 80 Days*, the globetrotter Phileas Fogg is described as a man whose "sole pastimes were reading the papers and playing whist." Fogg, arriving at the Reform Club before noon each day, peruses the newspapers until four in the afternoon. The narrative makes it clear he has some twenty papers to choose from. Reading the paper is a diversion that requires his full attention—he must hold a thought and follow a theme of exposition for long stretches of time.

This passage in the novel takes place in 1872. Read today, it says much more about the nature of the newspaper at the time than it does about Phileas Fogg—a century later, a man who spends half his day reading newspapers would probably be thought insane. Newspapers today contain the "news," but are hardly meant to be read, in the true sense of the word.

Along the way something has changed.

The media critic Neil Postman calls this change in the narrative structure of the news "a three-pronged attack on typography's definition of discourse, introducing on a large scale irrelevance, impotence, and incoherence." The news today, unlike that of Fogg's time, is "sensational, fragmented, impersonal ... to be noted with excitement, to be forgotten with dispatch" (Postman, 1985, pp. 65, 70).

Regardless of this change toward the daily news being fragmented and meaningless, as we read about Fogg's immersion in the paper, it is difficult to judge precisely why he must be so attentive to it. Is it because the language in the newspaper is difficult, or is it because he must keep track of a frail but coherent thread in an article? A second-hand account is not enough to infer how much of his time and mental effort is spent on each of the two tasks. To find an answer one must look at the actual news of Fogg's day and compare it with today's news.

This study is an effort to assess the overall linguistic complexity in average news articles printed in two major U.S. newspapers and a large news agency in the period 1900–2000, using cognitive theories about human sentence processing as a basis.

The focus of the study is on syntactic complexity and its impact on the time and effort a reader consumes in comprehending the articles in the material investigated. One of the goals of the study has been to understand how the style of newswriting in the United States has changed over the twentieth century in regard to linguistic complexity, and, in light of the data, to evaluate possible reasons as to why these changes have occurred.

The immediate data provided by the study will be supplemented with a look at external pressures on journalists to change their writing style from

time to time. Contemporary ideas about what kinds of sentences and constructions actually cause problems for human comprehension will also be discussed.

There are a variety of factors that contribute to making written text easy or hard to understand: the familiarity of the vocabulary used, the familiarity of the syntactic constructions used, the discourse context in which a sentence is introduced, the semantic plausibility of sentences, and the demands of the individual sentences on the cognitive sentence processing mechanism.

In corpus-based studies such as this one, the immediate end is to acquire quantitative evidence supporting or refuting a hypothesis about the material at hand. For this purpose, one of the most profitable foci of study is syntactic structure, which can be efficiently examined with the available technology. Other areas, such as semantic plausibility, offer little in the way of providing data that can be observed for trends, especially over long periods of time. Making accurate predictions about semantic plausibility is also very difficult. The world view and background of the archetypal 'average' reader may have changed—what was semantically plausible 100 years ago may seem outlandish today. There are few objective models that can track such changes and map them on an absolute scale.

Syntactic complexity, unlike vocabulary complexity, also does not vary much from person to person. A reader who is unfamiliar with the vocabulary in a given text can easily acquire fluency in the necessary words and attain a new reading level. The effect on complexity of syntactic structure is different from that of word choice in that there is very little that practice or learning will do to make diffuse sentences more readable. Many of the resources of the human sentence processing mechanism appear to have boundaries: what is a complex sentence for one reader, will most likely also appear that

way for others. Studies on sentence comprehension have shown remarkable uniformity between test subjects in this respect. This appears to be true in cross-linguistic settings as well: many syntactic phenomena that have counterparts in other languages are also identical in the way they complicate understanding.

As the focus is primarily on syntactic complexity, I have attempted to minimize other contributing factors of complexity in the news articles, such as semantic issues, vocabulary spread, and topic selection. In order to achieve this, the articles in the corpora have been narrowed down a single "type" of news article. The texts in the corpora are all relatively short (close to the average article length of daily newspapers), and have all been categorized as "general news." The assumption is that such articles should not include too exotic vocabulary or ideas, leaving in effect the syntactic choices of the writer as the main determiner of the articles' reading ease.

The aim in selecting sources for the study was to include articles from daily newspapers that primarily cater to a mass audience. The New York Times and The Washington Post are among the largest daily newspapers in the United States with a publication history that goes back over a hundred years. The Associated Press is the world's largest news agency, its articles appearing in nearly every newspaper published in the United States. The impact of these three widely read sources, which the material for this study is based on, will be discussed in chapter 2.

One of the objectives of the study is to see how the style of writing in these three sources reflects the intentions of the newspapers' internal guidelines for writing. Ever since the dawn of the "modern" U.S. newspaper in the early 1900s, most papers have aggressively tried to write in such a way as to be readable for as large a public as possible, thereby gaining an

edge over competitors in attracting potential subscribers. The newspapers have been in pursuit of producing easily readable material, and the details on how to reach this goal fill a good many pages in the writing guides that are written for the staff of the newspapers.

The first writing guides specifically targeted to journalists were published in the early 1900s. Before that the style of writing was largely dictated by a laissez-faire attitude, where the main business was getting a message across. From the first style guides all the way to in-house newsletters at end of the 20th century, a call for simple writing has been repeated in numerous ways. Chapter 3 looks at how newspaper staff and editors have been instructed to write.

In the 1920s, formulas that allowed for the calculation of "readability" were introduced. They sprang out of studies in education, and were part of an effort get the attention of writers by appealing to a "scientific" argument for simple writing. At the same time, they were the first efforts at quantitative determination of complexity. Readability was a new concept that was immediately put to use in the news business. The formulas seemed to reveal that the content of the average newspaper was unintelligible to a general reading public. Some claimed that reading and understanding a daily newspaper in the United States demanded a higher college education, or more. Rudolph Flesch, who developed a popular formula for calculating the readability of texts, was employed by The Associated Press to write a guide which would instruct its editors and staff in writing intelligible and interesting copy for the masses.

Many studies on the "readability" of newspapers have been published since the introduction of the readability formulas. Academic journals, such as Journalism Quarterly and the Journal of Mass Communication, have

regularly carried studies on various newspapers, magazines, and journals that have used readability measures to evaluate successes and shortcomings in the elusive goal of writing it straight and simple.

Given the impact on the news business of the notion of readability and the "readability movement," they will be studied fairly extensively, in chapter 4.

As the popularity of readability studies has diminished—or again retreated to the realm of educationists—newer ideas have surfaced about how humans actually take sentences apart and process them. Whereas readability studies focused on length of words and sentences, cognitive studies on sentence complexity have been founded on empirical evidence of sentence types whose syntax causes trouble to human understanding. Details of the cognitive studies are based on current linguistic theories and describe their results using syntactic models of sentences. Unlike the readability measures, the linguists' models do not treat words and syllables as discrete quantities that can be added up and put into a formula yielding a figure that says how understandable a sentence or a text fragment is.

The quantitative observations of syntactic complexity in this study largely follow The Dependency Locality Theory, first proposed by Gibson (2000). Many similar theories abound on complexity in human sentence processing, but the DLT has the advantage of being easily applied to computational tasks when using syntactically tagged corpora. The DLT is presented in chapter 5.

With the material and methods used, the study also offers an opportunity to compare the agreement of "readability" scores and results from the cognitive-based quantitative results. It has been held that even pure sentence length is as good an indicator as any for text complexity (Fry,

1988)—even though it is acknowledged that a short sentence may very well be bewilderingly complex, and a long sentence conversely easy to read. This and related ideas will also be evaluated in the results.

Chapter 2

The Newspapers and services

2.1 The New York Times & The Washington Post

The two newspapers included in this study, *The New York Times* and *The Washington Post*, are among the largest in the United States. They differ slightly in their intended audience, *The New York Times* is more of a national newspaper than *The Washington Post*, which holds a more regional base of readership.

The New York Times was established in 1851 as a "penny paper" which set out to avoid sensationalism and report the news in a more distant fashion. In 1901, it had a circulation of 100,000—well above average at the time. By 1921, it had reached 330,000 and in 1993, the daily circulation had grown to 1.2 million daily and 1.8 million on Sunday. By the end of the century, it was the third largest newspaper in the United States after the *Wall Street Journal* and *USA Today*, and had a circulation of 1.1 million.[1]

Founded in 1877, originally as an organ of the Democratic Party, *The Washington Post* slightly lags the *NYT* in circulation. It reported 162,000

[1]The Audit Bureau of Circulation, 2000

daily subscribers in 1933 to 1943. In 2000, it had a circulation of 762,000, making it the fifth largest U.S. daily.[2] It is considered the dominant newspaper in the U.S. capital.

2.2 The Associated Press

The Associated Press was founded in 1848 as a news gathering operation by newspapers on the U.S. East Coast. Its function was, and is, to provide its members and subscribers with news that otherwise would be difficult and expensive to obtain—particularly foreign and national news remote from the region of publication (Gramling, 1968). Nearly every newspaper in the United States received the *AP* newswire by 1990—some 5,000 radio and television stations and 1,700 newspapers in the U.S., as well as 8,500 subscribers internationally (Schwarzlose, 2002).[3]

Westley (1953) estimates that by the mid-20th century, wire service articles made up half of the printed content in the American newspaper. At the time, the *AP* supplied news to 60.3 percent of U.S. news outlets (Schwarzlose, 1979). The actual number of wire stories printed may in reality be much higher, though, mainly because of newspaper editors' reluctance to give credit to wire copy. In a study on the usage of wire services, Fenby (1986) notes a "tendency of many subscribers to use combined credits, or no credits at all."

Until about 1980, *The Associated Press* had a competitor in another wire service, *The United Press* (*UP* or *UPI*). The *UPI*'s presence started to diminish by the 1980s as overall newspaper circulation figures began to fall, and at the end of the 20th century, the *AP* was almost alone in supplying

[2] Ibid.
[3] Also, *http://www.ap.org*

wire news to U.S. newspapers (Schwarzlose, 2002).

The New York Times and *The Washington Post*, substantial newspapers themselves, also have their own newswire services—the *NYT* launched its service as early as World War I. Thus, the writing of the two newspaper sources used in this study also shows up outside their own pages (Fenby, 1986). In 1960 the *NYT* newswire had 60 clients. By the late 1980s, this number had increased to over five hundred, 350 of which were in the United States. *The Washington Post* service reached 363 at the same time, 199 of them in the United States (Fenby, 1986).

Chapter 3

The Establishment of Newswriting Style

Explicit writing guidelines directed at journalists began to surface in the late 1800s. At the time the majority of newspapers and agencies worked with small internal stylesheets and recommendations. Over time, however, writing guides and "stylebooks" became prominent and gained significant influence in newswriting.

3.1 Organization

The established style before the early 1900s was that of storytelling—a chronological account of the facts. The *AP* report on the assassination of Abraham Lincoln from April 14, 1865, is descriptive of this style. It opens:

> President Lincoln and wife, with other friends, this evening visited Ford's theatre, for the purpose of witnessing the performance of the "American Cousin."
>
> It was announced in the papers that General Grant would be present. But that gentleman took the late train of cars for New Jersey (Gramling, 1968, p. 56).

The unfolding of events thereafter proceeds in strictly chronological fashion, yet increasing in suspense: the theater was crowded; the audience delighted in the play; finally, the third act of the play begins and a sharp report is heard whereby confusion ensues, etc. About 100 words into the story it becomes clear that the president has been shot. Speculation over who the assailant might have been, eyewitness accounts or direct quotes are never given.

The instructions in early newswriting guidebooks focused on remodeling this particular story structure. Little is said about the details of exposition, such as sentence construction, vocabulary, and the like. The evolution that took place in the late 1800s was primarily toward a crude form of the "inverted pyramid," where the gist of the news is delivered first, then developing into more and more detail, in descending order of importance (Vos, 2002).

The motivation behind the inverted pyramid structure was based more on economical issues than on a concern for reader comfort. Hyde attributes the development to *AP* editors who wanted to simplify their duties of providing abridged stories to subscribers who paid for cheaper categories of service. The editors in the 1870s and 1880s simply wanted to avoid recasting stories into several lengths, preferring to get the job done by shearing off a suitable number of paragraphs at the end—"old-time newspapermen called the pattern 'the A.P. story'" (Hyde, 1952, p. 71).

Apart from the economical reasoning, the inverted pyramid would later be defended as a protector of objectivity with the argument that this standard structure prevented the writing of sentimental, lurid, and juicy stories in the spirit of "yellow journalism" (Vos, 2002).

Another explanation, given by Vos (2002), attributes the appearance of this structure to the widespread use of the telegraph in transmitting news stories. The rationale was that, if the telegraph lines were cut, the climax might remain untransmitted had it not been written in the inverted style, with the most important points first.

Gramling (1968, p. 103) holds that an 1899 *AP* story about a hurricane in Samoa is the first story in this new style, describing it as one that would "answer, in the first few lines, those five most pertinent questions—who, when, where why what." The first sentence in the story is nearly 100 words, but much more modern in its layout than the report on the murder of President Lincoln:

> The most violent and destructive hurricane ever known in the Southern Pacific passed over the Samoan Islands on the 16th and 17th of March, and as a result, a fleet of six warships and ten other vessels were ground to atoms on the coral reefs in the harbor, or thrown on the beach in front of the little city of Apia, and 142 officers and men of the American and German navies sleep forever under the reefs or lie buried in unmarked graves, thousands of miles from their native lands (Gramling, 1968, p. 103).

By the early 1900s, the majority of immediate news stories were written in the inverted pyramid style. One of the first guides to newswriting proclaimed that the format "tells its most thrilling content first and trusts to his [the reader's] interest to lead him on through the details that should logically precede the real news" (Hyde, 1912, p. 36).

Other guides followed suit, and the 'get-the-story-in-the-first-paragraph' style was touted in various books, and finally given its lasting name, the "inverted pyramid," probably in a 1934 textbook written by Carl Warren[1] (Vos, 2002).

[1]Warren, Carl N. (1934). *Modern News Reporting*, Harper & Bros. Publishers, New York.

3.2 Simplicity

The inverted pyramid and the conventionalized news form brought with it some simplicity since drawn out chronological accounts were more easily wrought in long sentences,[2] but the real call for the simple news story began in the 1940s with the introduction of readability formulas. The formulas had existed since the 1920s, but it took about two decades before they started to affect the newsroom.

The first such formula to come to the attention of the news business was the Flesch formula, which the author called a measure of "comprehension difficulty" (Flesch, 1943). Others followed suit, and a number of such readability calculations were suggested to the press in the 1940s and 1950s (Foulger, 1978).

The results that the formulas gave were not interpreted in a descriptive light as tools for evaluating different texts quantitatively. Rather, the propounders of the clarity yardsticks, along with newspaper editors, all saw a warning sign in the high complexity figures that the calculations yielded, and urged the press to write more plainly, "so they can be understood by the largest possible number of readers" (Campbell and Wolseley, 1961, p. 125). Flesch himself immediately jumped on the results emanated by the widespread application of his formulas and began advocating more 'shirt-sleeve English' in print (Campbell and Wolseley, 1961). Flesch delivered his revised formula to The Associated Press Flesch in 1948, which three years later printed his "AP Writing Handbook," where he advocated text that

[2] Journalists, however, also attributed simplicity to the need to save column space—something they would be constantly aware of in their trade: when the Washington Post announced in 1988 that it would change the spelling of *employee* to its current form from *employe*, an editor had an instant reply to comments that insinuated it was about time, too. He said that had the paper done the spelling change 10 years ago, 185 pages of print would have been wasted in printing the extra *e* (Bates, 1989, p. 60).

averaged 19 words per sentence, did not exceed 150 syllables per 100 words, and contained frequent use of what he called "human-interest" words.

The formulas, indeed, were not intended for academic investigation, but had a pragmatic purpose, by assuming a role of dictating how to write concisely:

> The common aim of all readability formulas has been to produce writing that would be simple and easy to read—especially for the less intellectual reader (Hyde, 1952, p. 126).

Soon, advice emphasizing simplicity was to be found in most guides to writing the news. Since the readability formulas were based on sentence length, word length, the average number of syllables, "human interest" words, etc., these aspects were tackled first, the slogan being the word "simple."

Despite the campaigning for a no-frills style, the campaigners often failed to exactly pin down the meaning of "simple," resorting to metaphor or vague advice. Reference to syntactic construction of sentences and how it contributed to complexity was not seen in the guides.

Jones (1949, pp. 23–24) called for a lowering of the "fog index," passing advice such as "write news leads that talk. Write the news like you would tell it"—obviously presuming that all journalists automatically use a narrower vocabulary and simpler constructions in speech than in premeditated writing.

Hyde (1952, p. 125) advises that a paragraph "must not exceed 10 or 12 lines of print," but that "writers may do much experimenting" with sentence length, even though readability formulas indicated that "no sentence should exceed 20 words."

Another suggestion was that a writer reverse the usual structure, not only of the story (as the inverted pyramid demanded) but of every single

sentence. A good writer "begins with the large idea and puts the qualifications later," i.e. avoids clause-initial adverbials, noun clauses, and the like (Hyde, 1952, p. 125).

Use of the passive voice has been equally well condemned:

> Occasionally, of course, news value dictates a passive; if you were writing about Mayor John, you'd want to lead with his name.
>
> In most cases, though, a passive is flabby, dropping the doer of a deed out of the picture. That's why officialese is addicted to the passive mode (Cappon, 1982, p. 26).

Not everyone agreed with the verdict passed out by readability formulas—that the writing was foggy. "Good writing is not a matter of mathematics or manipulation. There are no rules for it, except that it shall have feeling and individuality of impact," retorted Lester Markel, Sunday Editor of the *New York Times* (Campbell and Wolseley, 1961, p. 129).

Others acknowledged that a good portion of newsprint indeed was too challenging for the majority of the reading public, but took on the whole a didactic stance on the affair:

> But certain newspapers have made outstanding successes in mass circulation through a policy of 'shooting a little over the reader's head' and thus flattering him into reaching for higher standards (Hyde, 1952, p. 117).

Bush (1954) noted that more important stories were written using longer sentences, something later shown by Danielson and Bryan (1964) as well. This observation underlined the urgency of simplification: in the interest of fairness, democratic values, and keeping the public informed, the entire literate population had to be able to grasp important events. As Razik (1969, p. 324) put it: "if the newspaper is to be utilized to its greatest advantage as a means of mass communication it must be reviewed in the educational as well as the journalistic context."

If readability studies showed that the important news was cast in language impenetrable to the average reader, something was wrong and had to be rectified, mainly by telling those responsible—i.e. the journalists—to write in plain language.[3]

The urgent undertone was also partly caused by business-related concerns that the reader "turns to another story" if "a story lacks readability" or "looks dull and difficult to penetrate" (Campbell and Wolseley, 1961).

The view that simplicity was necessary because it sold papers and kept the public informed prevailed through the following decades. Particularly the 1970s, with the introduction of the computer, saw many academic readability studies—most of them using the Flesch formula—tailed with the ever-present advice to simplify. Hoskins (1973) showed that wire copy (AP and UPI) was at an 11-12th grade reading level for important events, and Burgoon et al. (1981) reached a similar conclusion that some national and international news was written above the reading level of about half the adult population, and hinted at the need to simplify to gain reader satisfaction.

Fundamentally, the writing advice to journalists remained the same throughout the rest of the century. By the 1980s and 1990s, it became more technically explicit, relying less on metaphor and the assumption that everyone would know what "simple" meant. Though unadorned writing had been urged in all writer's guides, toward the end of the century this simplicity was increasingly defined with reference to simple grammatical construction

[3]The task was not always easy, and journalists were difficult to persuade of the importance of simplification, as Bates (1989, p. 63) notes: "In the mid-1950s *The New York Times* tried to get reporters to write shorter, simpler sentences. Turner Catledge, the managing editor, told reporters to imagine that they were writing letters to a "curious but somewhat dumb younger brother." In *Winners and Sinners*, the paper's in-house newsletter, one of the editors instructed writers to limit themselves to one idea per sentence. Another *Times* editor, Lester Markel, replied by a memo: "I have read your edition of Winners & Sinners. It is a special edition. It interests me. No end.""

rather than sentence length or readability formulas.[4]

The *AP Handbook for International Correspondents* urges writers to "keep dependent clauses to a minimum," to "be bearish on adjectives and adverbs," with the warning that "if you expect your readers to alligator-wrestle your sentences, you'll find few volunteers" (Doelling, 1998, p. 42).

Simultaneously showing and telling, the guide proceeds to give a description of the specific nature of simplicity:

> Be civil to your sentences and allow them to follow the natural order of thought. Don't interrupt the flow with long opposites and relative clauses. That often puts the verb and predicate half a kilometer from the subject. It's confusing (Doelling, 1998, p. 42).

So, moderate sentence length and simple vocabulary were still on the list of essentials of good newswriting, but more focus was being put on simplicity of syntactic construction instead of vocabulary, which falls in line with cognitive research on ease of reading (see chapter 5). Cappon (1982, p. 31) writes about three things to avoid: "a gaggle of secondary detail," "abstract and general language," and "vagueness."

The pressure to simplify the style remains strong: almost every issue of the AP's internal newsletter *The Insider* published monthly between 1989 and 2002 remind its staff to cut down on complexity. The rationale is simple: "newspaper readership, people's attention span, and news holes have been shrinking" (Cappon, 1991). These in-house newsletters, similarly to those of other newspapers, reiterate "pleas for short, simple leads and sentences, for subject-verb-object constructions shorn of subordinate clauses and other

[4]By this time the readability formulas were handily available on most word processors. Instead of being an abstruse and time-consuming calculus, the press of a button could return a number that revealed the document's readability. This ease of readability calculation may have led journalists to once and for all dismiss such tools as vulgar. Incidentally, *The Associated Press* "Workbench" word processor used by all the agency's journalists as of 1998, includes no readability calculation function—although its spell checker does suggest alternatives to wordy constructions.

meanderings" (Cappon, 1990b).

A novelty in the writer's guides and in-house pamphlets toward the end of the century is that they reveal a belief in that the long-standing mission to achieve simplicity throughout the editorial chain has largely succeeded, or at least are more optimistic about the overall development—although the victory is not quite complete yet:

> We have made some progress. Spot checks here and there show a more acceptable average length—below 20, anyway. And more stories hit the 16–17 average, or came closer, than before. But it's clear that we still have a way to go (Cappon, 1990a).

> But let's face it, our report is in no imminent danger from excessive simplicity. When we run into problems, they usually come from the opposite direction: Clutter, involved phrasing, sentences that plod beyond their natural stops (Cappon, 1990b).

Chapter 4

Measuring Readability

Readability formulas are ways to numerically gauge the comprehensibility of text, or "ease of reading." They commonly give an arbitrary score, convertible to a "reading level" scale indicating what school grade level a text would be suitable for, or what some percentage (typically 70%) of students with a given number of years of schooling could read.

4.1 History of readability measurement

Klare (1963) traces the history of readability studies back 900 A.D. when the Jewish Talmudists who studied the Talmud laws began counting words and ideas in them to distinguish unusual senses of words from usual ones and then produce rough measures of reading ease.

The modern history of readability studies begins in the 1920s, when numerically precise "readability formulas" were introduced, largely as a result of the work of educational researchers (Chall, 1988).

The need for such an objective benchmark sprung from the demands of educationists and textbook writers who wanted to measure their material and thereby more accurately ascertain what audience would benefit most

from a given book. One of the first, if not *the* first, published study, Lively and Pressey (1923), is titled *A Method for Measuring the 'Vocabulary Burden' of Textbooks.*

Soon after the appearance of the first formulas, researchers proposed dozens of ways to measure readability and the demands a text put on its reader. Bruce and Rubin (1988) report hundreds of formulas being put forward between 1920 and 1980.

Later, books for adult readers were also targeted. The so-called "readability movement" (Klare, 1963) and the *Sub-committee on Readable Books of the Commission on the Library and Adult Education*, formed in 1925, called for more intelligible material for adults, and saw readability measurements as an objective way to prevent confusion and promote adult literacy.

The formulas were and are widely used in assigning textbooks for various grades and in preparing government texts. The societal stature of reading formulas was raised in the 1970s with the passing of several so-called plain language laws. Some U.S. states have introduced legislation that requires certain texts, such as insurance contracts and government regulations, to adhere to some measure of readability (Bruce and Rubin, 1988). In 2001, ten states in the U.S. along with many European countries have introduced plain language laws (Asprey, 2003). Often, the clarity of documents and writing prescribed by such legislation is tested by the Flesch reading ease formula, which is seen as a simple, objective, and successful way to measure complexity and comprehension (Bruce and Rubin, 1988).

4.2 Elements of formulas

Early studies on readability (starting in the 1800s) as well as the first published formulas focused on the frequency of difficult or rare words, and held

the view that a document's clarity could be measured by the number of words outside a list of "familiar words." Several formulas still in popular use today employ simple word lists as a basis for evaluating readership levels.

In the early studies, the basic elements in readability were considered to be: a) word familiarity, b) sentence length, c) word length in syllables. The first actual formula to give a numerical value to a document's readability was probably that by Kitson who in 1921 devised a measure based on sentence length and word length in syllables to study newspaper and magazine readability (Klare, 1963).

During the heyday of the readability formula, a number of factors were considered and proposed as fundamental measures of readability. Among these were:

- The number of different (unique) words (Vogel & Washburne, 1928)
- The number of prepositions or prepositional phrases (Gray & Leary, 1935)
- The number of common words, assessed through limited word lists, such as Thorndike's list in *The Teacher's Word Book*
- The ratio of "Anglo-Saxon" words to words of Greek and Latin origin (Lewerentz, 1930)
- The number of polysyllabic words (Johnson, 1930; Flesch 1948)
- The frequency of introduction of new "ideas" (McClusky, 1934)
- The "difficulty" of "ideas" (!) (McElroy, 1953)
- The ratio of "concrete" to "abstract" ideas (Morriss & Halverson, 1938)

- The average number of affixes in words (Flesch, 1943)
- The number of words per modifier (Bloomer, 1959)
- The number of "personal words" and personal pronouns (Flesch, 1948)
- The type of sentences used [simple, compound, compound-complex] (Vogel & Washburne, 1928)
- Indices based on "cloze" procedures, where a number of words are deleted from an existing text, after which the subjects' ability to fill the missing blanks in measured (Taylor, 1953)

4.2.1 Common formulas

Despite the abundance of ideas that have been put forth about readability formulas, the ones that have gained popularity only use a small number of observations. The three most popular formulas use primarily a combination of word length, sentence length in letters or syllables, and a predesigned "easy word" list.

Flesch

The Flesch formula uses only average word length (wl) in letters and sentence length (sl) to calculate the "reading ease" (R.E.) (Klare, 1963). In general, a randomly picked 100-word sample from a corpus is used for the formulas.

$$R.E. = 206.835 - 0.846wl - 1.015sl \tag{4.1}$$

The grade-level adjusted version (also called Flesch-Kincaid) only adjusts the constants in the first formula to scale the result to a "grade level," implying which the minimum education level for which a text is suitable.

$$R.E.G. = 11.8wl + 0.39sl - 15.59 \tag{4.2}$$

The Flesch Formula also includes a rarely-used "human interest" factor, which is calculated using the ratio of "personal words" (i.e. personal pronouns or proper names) to non-personal words, and the ratio of "personal sentences," i.e. sentences that contain personal words, to "non-personal" ones.

$$H.I. = 3.635pw + 0.314ps \tag{4.3}$$

Dale-Chall

The Dale-Chall formula uses the average sentence length in words (x_2), and the percentage of words outside an "easy" word list of 3,000 words (x_1). The result is the reading grade score of a pupil who would answer half of test questions on a passage correctly.

$$x_{c_{50}} = 0.1579x_1 + 0.496x_2 + 3.6365 \tag{4.4}$$

Gunning's Fog index

Gunning's Fog index uses the average sentence length and the percentage of words that have three or more syllables. Like the Flesch-Kincaid formula, it yields a number relating the readability to a "grade level."

$$0.4(\textit{average sentence length} + \textit{percentage of words of } 3 \textit{ or more syllables}) \tag{4.5}$$

4.2.2 Criticism of readability formulas

Criticism of readability formulas has broadly fallen into two categories—that the formulas themselves do not reflect actual human processing of text, and that changing sentences to adhere to the scores of a formula does not necessarily produce "readability."

It has been argued that formula-friendly writing leads to short sentences where the causality relation is often sacrificed, since splitting long sentences in two is naturally done at coordinating and subordinating conjunctions, leaving out words such as "because." Losing the connection between two causally related ideas makes for difficult understanding and poor recall, whereas longer sentences would add cohesion to connected ideas (Kintsch and Vipond, 1979). Among the educationists, Fry (1988, p. 81) concurs that a causality relation is lost, but calls the mere tweaking and chopping up of sentences to match a formula "cheating," adding that the formulas indeed work, but only "to judge the difficulty of a prose passage *after* it has been written" (original emphasis).

As mentioned earlier, most of the individual methods in fact only use two or three different counts to establish readability—as in Chall's or Flesch's formula. An example of extreme simplicity is the Lewerentz formula published in 1929, which simply counts the percentage of words beginning with *w,h,b* (easy words) and *i or e* (hard words) (Klare, 1963). The obvious problem with the method is that other aspects which are not counted may very well contribute strongly to making a given text nearly unreadable.

Bruce and Rubin (1988) note that non-countable phenomena—like syntactic complexity, the complexity of ideas, rhetorical structure and discourse variety, the number of items to remember during reading, and the number of inferences required—have been particularly underrepresented in the his-

tory of readability formulas. Others, such as Anderson and Davidson (1988) take the view that word difficulty and sentence length—the most commonly used factors in establishing readability—really have no bearing on the difficulty of a text, and call for assessment of readability to focus exclusively on linguistic aspects.

Randall (1988) notes that morphological complexity is a factor which has rarely been taken into account in readability studies. Her study, which tested both children and adults, reports a substantial but complex and still largely unpredictable relationship between morphological features of words and comprehension difficulty.

Baker et al. (1988) claim that all readability formulas work with an underlying, faulty assumption, derived from a traditional model of reading: that a passive reader "decodes" a text to obtain its meaning with no dependence on context or domains of knowledge.

Original studies in readability often did acknowledge (although somewhat superficially) the role of syntactic construction in comprehension, but commonly bypassed this, arguing that "more complicated sentences are generally longer than simple sentences"—which is why sentence length could be used just as well to measure complexity (Klare, 1963, p. 170). Similarly, Fry (1988, p. 80) argued that looking at syntactic details was unnecessary: the formulas "could measure grammatical constructions such as prepositional phrases and subjunctive clauses, but most of these measures correlate highly with average sentence length."[1]

[1]From an educationist perspective, this argument has a counterclaim in the correlation of T-unit length (roughly equivalent to independent clause length) and studies of syntactic maturity, established in Hunt (1964). Measuring the T-unit as opposed to sentence length is somewhat impervious to flux in punctuation practices, and results showed that some students wrote long or short sentences which did not at all correlate with their level of complexity. With this in mind, Fry's generalization that sentence length is usable just because it correlates with complexity does seem a bit broad-handed.

Probably the most common critique is that the measures themselves are devised ad hoc, and are largely unsupported by strict empirical evidence. Even though the measurements take into account features that seemingly contribute to reading difficulty, there is no demonstrable correlation between actual comprehension tests and reading difficulty scores. Also, cross-correlations between texts scored with different readability formulas have been found to have a large spread. As Kemper (1988, p. 152) points out: "these formulas cannot distinguish a well-structured text from a sequence of randomly ordered sentences."

Although simple formulas such as Lewerentz's (words that begin with *w,h,b,i* and *e*) have in fact worked to some degree—at least enough to embolden researchers to suggest them—this is obviously only because the occurrence of words beginning with *w,h,b,i* and *e* have correlated to some degree with complex writing, not because such words would be the cause of complexity. It is this use of a non-causal relationship to measure complexity that has been frowned upon by linguists, who have sought to define what causes complexity in sentence processing in the human mind, instead of looking at "surface" phenomena that possibly entail it under some circumstances.

4.3 Readability assessment based on cognitive results and sentence processing theory

Largely as a result of the critique outlined in the previous section, cognitive scientists have begun focusing on finding linguistic models to account for comprehension difficulties and have restricted their efforts to modeling complexity in terms of syntax. Many linguists and cognitive scientists in

the 1960s began expecting that theoretical models of syntax and semantics would offer a framework to precisely explain why certain texts or utterances were more easily comprehensible than others.

Chomsky (1965, p.15) notes that "it seems that the study of performance models incorporating generative grammars may be a fruitful study; furthermore, it is difficult to imagine any other basis on which a theory of performance might develop."[2]

Instead of focusing on the surface aspect of texts, researchers are turning more and more toward tying theories of complexity into linguistic models, preferably describing complexity at the point where grammatical and semantic information is mentally accessed. At the same time, much more reliance is put on experimental measurement of comprehension time and other related tasks.

Kemper's (1988) event chain modeling of text comprehension is an example of this type of assessment. In the model, what is calculated is the "inference load" on the reader—i.e. the amount and type of inferential processing the reader must perform when faced with a text. Kemper reports that the model has at least the equivalent predictive power of the Dale and Chall formula or the Flesch formula.

As this shift of paradigm has occurred, "readability" as the concept of measuring ease of reading through a study of surface linguistic features has been relegated mainly to the realm of education. Linguists, although they have tried to account for mechanisms that would explain the cognitive pro-

[2]It is noteworthy that while educationists were labeling readability with ever-increasing grade-levels (with the assumption that further schooling would increase text comprehension), theoretical linguists were quick to point out how certain types of complexity were not just a question of unfamiliarity or training, but most likely represented permanent limitations of the mind or the "language faculty." From this perspective, there was no amount of training or schooling that would ever bring a person to fluency in comprehending, say, multiply center-embedded structures.

cessing of sentences, have rarely proposed general formulas or other quantitative measures to explain complexity and sentence processing difficulty.

These sentence processing theories seek to explain—often with models that are construction and language-independent—what constitutes complexity. Some of the current theories lend themselves to quantitative assessment of the phenomena they describe, and are therefore the most likely candidates to serve as modern substitutes for what have traditionally been called "readability" studies. This development also shifts the definition of readability somewhat. Instead of measuring a vague idea of ease-of-reading, these theories can provide models to measure "cognitive load," be it syntactic, semantic, or memory-related.

Chapter 5

The Dependency Locality Theory

One of the recent models that has come out of an inquiry into how language is processed in the brain and what constitutes linguistic complexity is the Dependency Locality Theory, proposed by Gibson (2000). Like many other theories, such as the theory of Early Immediate Constituents (Hawkins, 1994), and Syntactic Prediction Locality Theory (Gibson, 1998), it is based on a view of sentence interpretation (or parsing) as a task that is primarily made difficult by demands on the computational and short-term memory resources of the brain, which are used for the on-line integration and accessing of a variety of information sources, primarily discourse referents. It relies heavily on the concept of locality, i.e. it works with an underlying hypothesis that there is a processing "cost" involved in constructing an interpretation of a sentence. This cost is primarily caused by an increase in distance between syntactic elements that refer to each other; words, or "discourse referents." The Dependency Locality Theory proposes a way to quantify this processing difficulty in a given sentence. This processing cost calculation has been sup-

ported by a number of empirical studies where reading times have correlated strongly with predictions made by the DLT.

The bulk of the processing cost in the DLT theory relates to the integration difficulty when a reader encounters a finite verb, and must link it to the (generally) preceding subject noun head. Increasing the distance between a verb and its arguments complicates integration at the verb. This is true of other kinds of integration as well—attaching prepositional phrases to nouns and verbs and attaching verbal phrases, etc.—though the DLT specifically addresses only the issue of attachment at finite verbs.

5.1 Types of syntactic complexity

5.1.1 Complexity caused by ambiguity

Ambiguous sentences have long been known to cause processing trouble because of the large number of competing interpretations that a reader has to choose from; for example:

> (1) The board approved its acquisition by Royal $Trustco^{1}$ Ltd. of $Toronto^{2}$ for $\$27^{3}$ a share at its monthly $meeting^{4}$ (Manning and Schütze, 2000).

The above sentence has at least 5 readings, because the numbered prepositional phrases **by Royal Trustco Ltd.**, **of Toronto**, **for $27**, and **at its monthly meeting** can be attached in a variety of ways. Many of these attachment options are transparent to the reader and are solved largely by semantic means.

The number of readings through the choice of attachment of prepositional phrases increases as the combinatorial series of Catalan numbers

(Church and Patil, 1982). For every new PP, the number of readings increase to: 1, 2, 5, 14, 42, 132, 469, 1430, etc.[1]

Accordingly, when processing PPs there are two factors at play that must be distinguished: ambiguity and distance of attachment. The level of ambiguity of a sentence is a very subtle concept, not easily quantified, whereas the distance of attachment to the noun or verb that the PP modifies can easily be observed.

The processing of ambiguous sentences is subject to much variability depending on the background and knowledge of the readers.[2] To find the intended meaning in the following two sentences requires intricate judgment about extralinguistic plausibility:

(2) I examined the man with a stethoscope.
(3) I examined the man with a broken leg.

5.1.2 Complexity in unambiguous sentences

The DLT has grown out of studies where the goal is to account for complexity primarily in unambiguous sentences and in their processing.

Unambiguous sentences can cause complexity in a different way than ambiguous ones. As they become harder to read, they seem to reflect the difficulty of the human processing mechanism to keep up with the job of parsing, rather than an effort to select the most plausible reading.

Nesting

These are many cases where sentence processing energy is mainly spent on integrating discourse referents with each other, not on selecting between

[1] $C_n = \frac{2n!}{n!(n+1)!}$

[2] This is why computer parsers have such difficulty with PP attachment—the task requires very subtle decisions that involve knowledge about the world, something not easy to handle computationally.

several likely parses and deciding which is the most feasible one. A typical group of these sentence types that are unambiguous, yet notoriously difficult to process are nested, or center-embedded structures. Miller and Chomsky (1963) noted that this type of structure was much more difficult to parse than a right-branching or left-branching structure, regardless of syntactic ambiguity—as is seen in the sentences below, where the complexity of the center-embedded structure goes beyond the comprehensible and overloads the sentence processing mechanism:

> (4) Right branching: The dog that chased the cat that ate the rat barked.
> (5) Nested: This is the malt that the rat that the cat that the dog worried killed ate (Yngve, 1960)

Garden pathing

Center-embedding also differs from the near-ambiguity of what are called garden-path structures, such as the classic example:

> (6) The horse raced past the barn door fell

The garden-path structure is immediately resolvable once the reader knows the intended meaning—there is really no syntactic ambiguity present. But the reader who is performing a moment-by-moment integration of the constituents is tricked down a dead-end parse which does not become evident until the last word is reached. The garden-path sentence, as opposed to the earlier nested example, is obvious on subsequent readings, whereas the center-embedded example is not.

5.2 Integration cost and storage cost in the DLT

The DLT makes predictions about different types of syntactic complexity using two components. One is an "integration cost" component that takes into account the work done when joining syntactic elements, as in attaching **rat** to **ate** in (5). The other is a component for the "storage cost" associated with keeping track of the sentence structure during reading.

5.2.1 Storage cost

The memory storage cost varies during the reading of a sentence and increases when deep inside a nesting. The storage cost is calculated by observing the number of constituents needed to complete the sentence grammatically. So for example the sentence

> (7) The reporter who the senator attacked disliked the editor (Gibson, 2000)

has its maximum storage cost requirement at **the senator**, where 4 words would be needed to complete the sentence grammatically. At the point **the** (that begins the sentence) two syntactic heads would be needed to complete a grammatical sentence.

Because storage cost is a moment-to-moment observation that reveals how much processing is going on at a single point in reading a sentence, it will not be included in the quantitative methods of this study. It is worth noting, however, that the immediate outcome of the theory is that the processing occurs at specific points when reading a sentence, not as an overall effort, and that, the deeper the nesting in a sentence, the more energy is continually being spent while inside the nesting.

5.2.2 Integration cost

The structural integration cost is present when connecting a syntactic head h_2 to an existing syntactic head h_1. It increases in proportion to the number of new discourse referents that intervene between the two elements.

Gibson (2000) proposes a simplified model for calculating the amount of energy spent for integration, dividing the cost into a cost for introducing new referents, and a cost for integrating referents:

1. One energy unit (EU) is spent for each new discourse referent that is introduced. A discourse referent is either a head noun or a verb.

2. The cost of integration of a new head h is calculated by the number of new discourse referents introduced between h and all heads h' that are dependent on h. Simplifying, 1 EU is spent for each intervening discourse referent between heads h and h'. Verbs are also counted as discourse referents. Pronouns are excluded, as they refer to something already present in the discourse (see table 5.1 on page 36).

5.2.3 Semantic plausibility

The DLT acknowledges that semantic considerations affect the processing demands and can make the meaning of an otherwise complex sentence easy to construe, as in:

> (8) The vase that the maid that the agency hired dropped on the floor broke into a hundred pieces (Gibson, 2000)

In (8), which is relatively easy to follow, the distinctive semantic classes of the NPs aid the integration process, in contrast to others which are similar in structure but where the NPs and their respective VPs are not so easily

Cost type	The	reporter	who	the	photographer	sent	to	the	editor	hoped	for	a	good	story
New referent	0	1	0	0	1	1	0	0	1	1	0	0	0	1
Integration	0	0	0	0	0	2	0	0	0	3	0	0	0	0
Cost type	The	reporter	who	sent	the	photographer	to	the	editor	hoped	for	a	good	story
New referent	0	1	0	1	0	1	0	0	1	1	0	0	0	1
Integration	0	0	0	0	0	0	0	0	0	3	0	0	0	0

Table 5.1: *Illustration of the costs of integrating discourse referents to structural heads (after Gibson (2000)). In the first example, structural integration begins at the verb* sent. *The object-extraction are higher than for subject extraction, where the verb* sent *is a local integration, performed at no cost according to the DLT model. Gibson et al. (2004) and have shown this model to agree with actual reading duration in self-paced reading time experiments.*

connected. However, this semantic distinction is something that is not easily quantified, especially when working with computational tools. Therefore, no distinction of semantic plausibility will be made in this study.

5.3 Examples of the DLT at work

5.3.1 Pronouns

The DLT accounts for a number of different effects in sentence processing, quantifying why some very similar structures demand varying amounts of processing energy.

Certain types of embedded sentences are known to be easier to process than others. Particularly those laden with pronouns tend to be easier, as in:

> (9) A book that some Italian that I have never heard of wrote will be published soon by MIT Press. (Frank, 1992)

This example sentence, regardless of its nesting, is not as complex as others that are structurally similar, for example:

> (10) The reporter who the senator who John met attacked disliked the editor.

The main difference is that the prior example contains a pronoun in the nesting, which makes it much easier to process. The DLT automatically accounts for this when calculating integration costs, because pronouns are not counted as "new discourse referents."

5.3.2 Calculating complexity

Here are some examples of the increasing difficulty for the sentence processing mechanism to handle long-distance dependencies.

(11) The professor copied the article.
(12) The professor who advised the student copied the article.
(13) The professor who collaborated with the scientist who advised the student copied the article.

The increasing demands, or "integration cost," of processing sentences 11–13 are primarily dependent on the number and duration of incomplete syntactic heads starting with **The student**. In (12), two other referents are introduced before resolving the first NP by the verb **copied**. Following the DLT model, sentences 11, 12, and 13 receive total integration costs of 0, 2, and 4—(11) has no intervening discourse referents between **professor** and **copied**, (12) has two, and (13) has four.[3]

5.4 DLT use in this study

Since ordinary corpora, newswriting for popular digestion in particular, is unlikely to exhibit the kind of extreme processing difficulty seen in examples favored by cognitive research, the idea here is to look at more subtle variations in the distance of syntactic nodes that need to be integrated throughout the processing of a sentence. Special attention is paid to the number of incomplete dependencies that a parsing mechanism must keep track of during processing. Since this is a corpus-based study, contributions to sentence complexity due to overlapping semantic classes which are difficult to assess without case-by-case human evaluation will be discussed minimally, leaving the primary focus on syntactic distance.

In effect, what the study is looking for is the frequency and usage of constructions such as:

[3]Object-extracted relative clauses, such as *The student who the professor advised graduated* are more demanding. This is because the **who** is attached to an empty catory *e* following **advised**, bringing the integration cost just within the relative clause to two.

> (14) **The ordinances** passed yesterday by the Cabinet Council, authorizing a guarantee of the principal and interest of an issue of 10,000,000 yen ($4,890,000) debentures for the purpose of expediting work on the Seoul-Pusan Railway, and which also provided for all possible military expenses for the protection of the railway and other interests, also **authorize** the Government to utilize 50,000,000 yen, ($24,450,000) the proceeds of the Chinese war indemnity, which hitherto has been devoted to educational and other purposes, as a war fund. (AP Dec 29, 1903).

Here, there is a substantial distance between the subject (**ordinances**) and the main verb (**authorize**). Several new discourse referents are introduced—eleven NPs in all—while demanding that the reader keep track of the incomplete dependency started by **The ordinances**.

5.4.1 Nonlinearity of phenomena

It must be noted that of the different kinds of integration that may occur during the processing of a sentence, not all are comparable in their expenditure of processing energy.

It seems likely that a single long integration may not be as complex as several short ones occurring frequently, although the sum total in energy units would be the same.

Secondly, there is little evidence on how different types of integration relate to each other. Obviously, the introduction of new clauses that interpose either a noun and a main verb, or the main verb and the complement, require the most processing energy. But it is unclear how this processing relates to other types of integration, for instance, attaching prepositional phrases to noun or verbs, or attaching verbal phrases. Also, connecting the verb and the complement in a verbal phrase is probably not as energy-consuming as when dealing with a finite verb.

It has been observed that if only nouns interfere between a subject and

the main verb (such as several appositive phrases), complexity is primarily caused by memory decay—the tendency to forget the main subject when the verb finally arrives. This decay is much faster if any intervening noun is linked to a verb, such as in a relative clause—which always contains a finite verb.

These different flavors of integration may not even be comparable. In this study, a range of phenomena (subj-verb attachment, verb-obj attachment, PP attachment) have therefore been investigated and the results are reported separately, although the counting process is the same: counting the number of intervening new discourse referents between the syntactic heads.

5.4.2 Computational implementation

One advantage with the DLT when working with corpora is the possibility of automating the process of calculating integration costs—presuming the corpus to be studied is parsed in enough detail to allow for these calculations. This naturally depends on the format in which texts are tagged and parsed (simple part-of-speech tagging will not suffice). The tool used in this study, Connexor's FDG parser (Tapanainen and Järvinen, 1997), allows for automatic analysis of some types of integration cost, primarily integration and new referent cost.

5.5 The DLT vs. readability

Studies in linguistic complexity and sentence comprehension focus on the question of complexity in much more detail than readability studies—phenomena are separated and broken down into the smallest elements that can be isolated and studied in empirical contexts. In the DLT, not all such elements are studied in detail. Particularly semantic factors are only included insofar

as noun phrases are concerned, and then only to provide a model for differentiating pronoun types, proper names, and definite NPs inasmuch as their attachments contribute to complexity.

Despite the somewhat different approaches of readability measurement and linguistic complexity theories, one of the ultimate goals of the DLT is to provide a theory of comprehension time and "intuitive" complexity (Gibson, 2000). While readability measures, in contrast to the DLT, also provide metrics for the impact of vocabulary on comprehension, the DLT says nothing about the role of vocabulary.

In light of the scope of the DLT, this study will focus on syntactic complexity measurement and vocabulary quality, or dispersion, as two separate factors in the assessment of comprehensibility of text.

Chapter 6

Studies in Newspaper Readability

A large volume of studies concerning newspaper style, especially in the United States, has been published since the dawn of readability formulas. Much of the scholarship has been done with a somewhat prescriptivist slant, using the results to urge newspapers to write more simply. In one study that was purely quantitative, the author added after the results: "If the newspaper is to be utilized to its greatest advantage as a means of mass communication it must be reviewed in the educational as well as the journalistic context" (Razik, 1969).

Since Flesch's 1943 readability studies were published, the majority of newspaper research has reported its results in this format. Sometimes the Dale-Chall formula is used. Average sentence lengths are customarily given as well.

Very few reports of changes in complexity or "readability" over time are found in the literature. Most studies have compared the readability of different sources during a specific point in time, with very little categori-

cal analysis about what kind of sentences and constructions contribute to making articles difficult to read.

Fowler (1978), using the Flesch-Kincaid scale, compared the readability of newspapers and novels during the years 1904, 1933, and 1965.[1] Fowler found that newspapers were significantly more difficult to read in 1933 than in 1904, but then much easier to read in 1965. Sentence length was 28.47, 28.82, and 21.82 during the three years, respectively.

Klare (1963) quotes several studies done in the early fifties, revealing a trend at the time toward simplification compared with earlier material. The average sentence length in newspapers had settled around 23 words, wire story leads being slightly longer.

Hoskins (1973), using the Flesch formula, compared the readability of *AP* the *UPI* wire stories from 1972, and found that the majority of both news wires were on a 13th to 16th grade reading level. The *UPI* copy was slightly more difficult. The average sentence length for *AP* copy was 23 words.

Razik (1969) tested a sample of newspapers from 50 cities (using the Dale-Chall formula) and found that national and international news were at an 11–12 grade level in all newspapers, while the majority of article types in non-metropolitan newspapers were at an 11–12 grade reading level, and 9–10 in metropolitan newspapers.

Jung and Jo (2001), using the Flesch-Kincaid formula, investigated business news sections and found that *The New York Times* contained exceptionally difficult material compared with other newspapers, such as *USA Today*, *The Wall Street Journal*, and electronic publications.

Danielson and Bryan (1964) who studied the readability of different cat-

[1]The time frame of Fowler's study matches the first three time periods of this one.

egories of wire stories, reported that "soft" stories were easier to read than "hard" stories[2].

Lead length in the 1990s has been studied by Stone 2000. "Prestige" dailies, such as *The New York Times*, were found to use significantly longer leads in their stories compared with non-prestige newspapers. Only *USA Today* adhered to using short leads in active voice, as per the recommendations of newspaper style guides. The prestige papers' leads averaged 27.6 words, and non-prestige dailies 22.3.

The Associated Press, doing internal checks, reported on reaching a lead length of less than 20 words in 1991 (Cappon, 1990a).

[2]Other research has noted the same result. The material in this study is "hard," given the selection of short spot news articles.

Chapter 7

Approach

7.1 Material

The material used in this study were articles from two major newspapers in the United States—*The Washington Post* and *The New York Times*—as well as from *The Associated Press* newswire.[1]

Articles were selected from random dates from four decades during the period 1900—2000 from each of the three sources, though the selection was weighted toward the middle of each decade. For every decade, samples of at least 10,000 words were gathered. All the articles were what are called "spot news"—i.e. news of immediate impact—a straight, stylistically unembellished report of some occurrence or event without any analysis. All of the articles were less than 300 words, as longer articles were judged more likely to deviate from the style of immediate reporting. All the articles were also written by the paper's or agency's own staff—stories where the credit line mentioned another source, or a combination of sources, were ig-

[1]The articles were gathered from *ProQuest Inc.*'s Historical Newspapers archive, which offers the complete *New York Times* and *Washington Post* papers as digital images. The images were processed with Optical Character Recognition software and proofread manually.

nored. "Pickup" stories, that mostly attribute information to secondary sources, were also removed from the corpus during proofreading (whenever they were spotted). Stories where more than 50% of paragraphs were in the form of direct quotes were also omitted.

Another requirement of the coding was that no article carry a byline, since it is commonly the case that when a writer is granted a byline, instead of sticking to the established "in-house" style, the writer will give personal style freer play.[2]

Finally, the material was pruned and divided into twelve subcorpora by decade and source (i.e. four for each of the three sources) representing articles from 1900–1910, 1930–1940, 1960–1970, and 1990–2000. The articles contain a mixture of international, national, and metropolitan news stories, approximating in proportion the division of material within the individual newspapers. For every random date, all articles that fit the above coding criteria were taken from all three sections; the selection should thus reflect the actual sizes of the sections in the papers. There is some disparity in the division of material. The Washington Post has a fairly extensive metropolitan section,[3] but less national and international news of their own pen. The *NYT* has a larger national and international desk than the *WSP*. The *NYT* also prints more of its own reporting than the *WSP*, especially as regards international and national news. The *AP* does not carry "metropolitan news," rather, all of its output is classified as either national or international. The

[2]By the 1990s, The Washington Post had introduced a practice of almost always bylining their articles, no matter how short, and the requirement was overlooked for that subcorpus.

[3]However, it should be kept in mind that labeling a story as "local" is possible even when the news originates from afar. "On a San Fernando Valley radio station, where management had ordered that every newscast open with a local story, one newscast began: "Two high-speed trains collided today between Tokyo and Osaka, Japan. There were 123 people killed and several hundred have been injured. But there were no Valley residents on board"" (Bates, 1989, p. 13).

AP material was selected from articles that had actually been published in one of three major U.S. newspapers—the *NYT*, the *WSP*, or the *Los Angeles Times*. Since the *AP* may file stories on its wire that never get published, the *AP* material for this study was taken from the three newspapers to assure that the articles have been seen in print and reflect actual usage.[4]

The material was tagged with Connexor's Machinese Syntax (previously FDG) parser[5] (Tapanainen and Järvinen, 1997), (Tapanainen, 1999). The Machinese Syntax tagger is a syntactic parser producing functional trees that present morphological information for word-form tokens and functional dependencies representing relational information in sentences. The output of the parser was then used as input for a number of programs written for this study, which—based on the syntactic relations, the functional tags, and morphological tags—calculated complexity counts and breakdowns of all sentences of the corpora.

Here is the output from the parser for an example sentence:

```
1       The     the     det:>2  @DN> %>N DET
2       ball    ball    subj:>10        @SUBJ %NH N NOM SG
3       given   give    mod:>2  @-FMAINV %VP EN
4       to      to      ha:>3   @ADVL %EH PREP
5       the     the     det:>7  @DN> %>N DET
6       Wanderers'      wanderer        attr:>7 @A> %>N N GEN PL +name
7       Club    club    pcomp:>4        @<P %NH N NOM +loc
8       last    last    det:>9  @DN> %>N DET +temp
9       night   night   tmp:>3  @ADVL %EH N NOM SG +temp
10      was     be      main:>0 @+FMAINV %VA V PAST
11      the     the     det:>13 @DN> %>N DET
12      most    much    ad:>13  @AD-A> %E> ADV SUP
13      magnificent     magnificent     comp:>10        @PCOMPL-S %NH A ABS
14      held    hold    mod:>13 @-FMAINV %VP EN
15      in      in      loc:>14 @ADVL %EH PREP
16      Havana  havana  pcomp:>15       @<P %NH N NOM SG +name
```

[4] A pilot study based on the same material with random *AP* wire copy from the 1970s and 1990s vs. AP copy that was published in at least one newspaper revealed that the generic AP copy is somewhat simpler, has shorter sentences, and spans a narrower stylistic range than those articles that actually get published in major newspapers. This may reflect a selection process among newspaper editors where they are more likely to pick wire stories for print that are slightly more complex (i.e. a question of prestige). Or it may be that stories of higher news value—those that are more likely to end up in print—are written in a more complex style.

[5] http://www.connexor.com

```
17      within  within  ha:>14  @ADVL %EH PREP
18      the     the     det:>19 @DN> %>N DET
19      memory  memory  pcomp:>17       @<P %NH N NOM SG
20      of      of      mod:>19 @<NOM-OF %N< PREP
21      the     the     det:>23 @DN> %>N DET
22      oldest  old     attr:>23        @A> %>N A SUP
23      leaders leader  pcomp:>20       @<P %NH N NOM PL
24      in      in      mod:>23 @<NOM %N< PREP
25      society society pcomp:>24       @<P %NH N NOM SG
26      .       .
27      <p>     <p>
```

The ability to count the distance and intervening elements between dependencies that contribute to complexity is the most important aspect of the tagging. In the sentence above, **ball** is tagged as a subject nominal head, corresponding to the main verb **was**. Between them are two new discourse referents (by the method of counting used here, see chapter 5): **given**, and **Wanderers' Club**. This yields a subject-verb complexity count of 2, actually far above the average complexity of sentences in any of the corpora.[6] Prepositional phrase attachment is revealed similarly in column 4 of the output—words numbered 4, 15, 17, 20, and 24 have their attachments correctly marked. Phrases occurring at words 20 and 24 modify nouns, the rest are adverbials.

7.2 The computation of complexity

Thusly, the integration cost of sentences was performed in three different areas, in keeping with the simplified model given in (Gibson, 2000) (see chapter 5). Separate counts were given for the integration costs of nominal heads to VPs, the cost of attaching PPs to nouns or verbs, and the attachment of verbs to objects. Adverb attachments were counted together with the attachment of PPs. Here is an example sentence and its cost calculation:

[6]The actual average S-V integration cost per sentence is about 0.7; see figs 8.1, 8.2, and 8.3.

> John Edwards, 44, making his first bid for public office, won the Democratic primary with 51 percent.

Here, attaching the subject, *John Edwards*, to *won*, is given an integration cost of three, because of the three intervening discourse elements—*making, bid, public office*. The apposition, *44*, is not counted. Attaching *for public office* to *bid* is done at no cost since no elements intervene. The phrase, *with 51 percent*, costs one PP attachment unit on account of the intervening, *the Democratic primary*.

7.3 Readability calculations

For comparability with previous studies about newspaper language, readability measures were calculated for all the corpora.[7] The Flesch, Flesch-Kincaid, and Fog measures are included in the results.

7.4 Accuracy

The tagging accuracy of the syntactic and part-of-speech tags was checked by manually correcting 50 tagged sentences from each of the 12 subcorpora, and then comparing the results with the machine tagging. The tagging accuracy varied between 96.6% and 97.1% agreement between the manually corrected and the machine tagged versions. Not all the tagging errors had any bearing on the subsequent complexity calculations.

All the sub-corpora remained in the same ranking of complexity in the manually tagged versions as in the computer tagged ones, i.e. the trend from the early 1900s to the 1990s remained the same even in the smaller

[7] These were calculated with a slightly modified (to handle sentence length counts better) version of the Perl tool Lingua::EN::Fathom, written by Kim Ryan, available at *http://search.cpan.org/ kimryan/Lingua-EN-Fathom-1.08/*

sample.

7.4.1 Notes on the tagging

There are two main types of features of importance to this study that machine taggers apparently have problems with: extremely long subject head noun and main verb separation, and attachment of prepositional phrases.

Very long subject-verb integration are often missed, and sometimes inaccurately tagged so that the subject noun is linked too early, often to a verb within a relative clause. The tagging is accurate up to a separation of 30 words or so, after which the results vary. In a preliminary study based on only the 1900–1910 and the 1990–2000 AP material, all of the sentences were checked by hand, and the earlier material did contain two extremely long separations of subject NP and VP (more than 50 words), not recognized by the tagger, whereas the 1990–2000 material did not have a single one of more than 30 words.

Attachment of other elements sometimes causes similar trouble, particularly if the correct parsing would require world knowledge, or knowledge of semantic aspects of a sentence, as in:

```
1 The the det:>2 @DN> %>N DET
2 shootings shooting @NH %NH N NOM PL
3 in in mod:>2 @<NOM %N< PREP
4 Atlanta atlanta pcomp:>3 @<P %NH N NOM SG +loc
5 yesterday yesterday tmp:>7 @ADVL %EH N NOM SG +temp
6 were be v-ch:>7 @+FAUXV %AUX V PAST PL
7 reported report main:>0 @-FMAINV %VP EN
8 not not neg:>9 @ADVL %EH NEG-PART
9 long long dur:>7 @ADVL %EH ADV
10 after after pm:>12 @CS %CS CS
11 they they subj:>12 @SUBJ %NH PRON PERS NOM PL3
12 began begin @+FMAINV %VA V PAST
```

Here, "yesterday" should attach to "shootings," not to "reported." The correct tagging would add to the cost of integration as the attachment of the adverb would cross over the prepositional complement NP "Atlanta."[8]

[8]This tagging example and awkward placement of the adverb also exemplifies the daily

	AP				NYT				WSP			
Decade	**00s**	**30s**	**60s**	**90s**	**00s**	**30s**	**60s**	**90s**	**00s**	**30s**	**60s**	**90s**
Articles	100	100	71	74	102	79	74	65	71	90	74	62
Words	10549	10455	10105	11447	10523	11023	10063	9993	10167	11271	11575	10122
Sentences	461	510	552	588	532	504	516	449	477	534	590	527
Avg. Length	22.88	20.50	18.30	19.46	19.78	21.88	19.50	22.26	21.31	21.11	19.62	19.21
Fog	17.59	14.92	14.78	14.99	14.34	16.03	15.37	16.28	15.60	16.21	14.79	15.20
Flesch	36.40	45.38	41.39	42.30	48.78	42.53	40.28	38.83	45.00	39.77	43.11	41.81
Flesch-Kincaid	13.87	12.02	12.04	12.20	11.37	12.77	12.49	13.37	12.28	12.96	12.12	12.20

Table 7.1: *The basic material of the study, including the average sentence length and readability scores.*

However, these errors also seem to have evened out as the manually tagged sub-corpora did rank in the same order of complexity in PP attachment as the machine-tagged ones.

problem of journalists grappling with where to drop the mandatory "time element" in the beginning of the story without making the sentence cumbersome to read. Most writers would probably prefer a prenominal modifier in this example to bring down the integration cost: *Yesterday's shootings in Atlanta....* On the other hand, *The shootings yesterday in Atlanta...*, which would also be economical and not cross over other NP's, sounds awkward, and many stylebooks caution against such a placement of the time element.

Chapter 8

Results

I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind. —Lord Kelvin

The overall nature of the results shows that subject-verb integration is by far the most significant factor in the accumulation of complexity. Integration of other elements—the verb and the complement, prepositional phrases, etc.—contribute with only a fraction to the overall processing requirements of sentences.

8.1 Complexity induced by subject head noun-verb attachment

All three subcorpora show a trend of moving toward more frequent interruption in the nominal head–verb chain.

In the *AP* 1900–1910, 28.60 percent of all subject noun heads were immediately followed by the verb. In *AP* 1990–2000, this figure has come down

to 24 percent. The *NYT* and *WSP* show a similar pattern in 1900–1910 vs. 1990–2000: 26.56 percent to 20.71 percent (*NYT*), and 26.01 percent to 20.28 (*WSP*). See table 8.1 on page 60 for the breakdown between the three corpora.

On the other hand, the processing cost of connecting a verb to a subject noun head has decreased over time (with the exception of the 1990–2000 subcorpora, where the trend is reversing). The same trend is evident when counting simply the number of words interposed between a subject head noun and a verb. This decreases steadily from 2.40 words in 1900–1910 to 2.10 words in 1990–2000 in the *AP* material, 2.43–2.21 in the *NYT*, and 2.27–2.06 in the *WSP*.

The combination of this data indicates a distinction between two major trends. Sentences are becoming shorter, subject head nouns are on the average drifting closer to the main verb over time—as style guides have advised—but, at the same time, sentences are more often broken up by interposing short clauses and phrases between the subject noun and verb, causing frequent minor fragmentation. This breakup of the subject and the verb actually increases the overall complexity toward the 1990s, as compared with the 1960s, which is a low point in complexity in the *NYT* and the *AP*.

An example story in the 1990s style reads:

> The wife of Gen. Colin L. Powell says in a forthcoming television interview that it would be dangerous for her husband to run for President because a black candidate would probably become a target for "crazy people."
>
> Alma Powell, speaking in an interview for the ABC News program "20/20" on Friday, said of her husband: "He would probably be at much more risk than any other candidate because of being a black man in this society. A lot of crazy people out there."
>
> But if General Powell, who is retired, does decide to run, Mrs. Powell added, "I will adjust."
>
> General Powell, a former Chairman of the Joint Chiefs of Staff, said his wife's feelings could be decisive. (*AP*)

This four paragraph story is very descriptive of the 1990s style, and shows where the complexity comes from. The participial phrase, *speaking in an interview for the ABC News program "20/20"*, introduces four new discourse elements before the main verb *said*. In paragraph three, the subordinate clause, *who is retired*, adds one referent. And in the final paragraph, the often used appositive adds two. The last sentence is 18 words long—shorter than the average—but more complex, though still a very mild example.

This can be contrasted with the "1960s style" in the corpus, where stories often begin without any detail interposed at all:

> A nationwide department store chain announced today that it would extend credit to people on welfare.
>
> Montgomery Ward and the National Welfare Rights Organization said that 3,000 welfare recipients holding membership in the organization would each receive up to $100 in credit under a one-year pilot program.

This is clearly a different type of rhetorical layout. The first sentence provides only generalities; all the details come in the second paragraph. This style has a significant impact on the complexity in sentence processing—if one were to squeeze the details of the second paragraph in this story into the first, avoiding syntactic complexity would become a cumbersome task in itself.

This interposing of items between subject and verb is on the rise in the 1990s in both the *AP* and the *NYT* corpora. In the *AP* corpus, the DLT cost of head noun-verb integration is 3.78 per 100 words processed[1] in 1900–1910, then 3.34 (1930–1930) (t=9.173), 2.94 (1960–1970) (t=10.141) and finally up to 3.12 (t=-3.296) in the 1990–2000 subcorpus.[2] The *NYT*

[1] The cost per 100 words was used as the primary average measure of processing cost. Average processing cost per sentence was also calculated at times to check if sentence length had any bearing on complexity.

[2] $p < .001$

follows a similar development of peaking at 1900–1910 at 3.76, then moving down to 2.95 (1930–1940) (t=3.296), 2.58 (1960–1970) (t=11.342), and up again to 3.20 in 1990–2000 (t=-17.355).[3]

The *WSP* corpus shows no increase toward the 1990s. Instead, the trend is constantly toward simpler attachments and less distance between nominal heads and verbs, the subject-verb integration cost being: 3.11, 3.13, 3.10, 2.78—in 1900s,1930s,1960s, and 1990s (t=7.263, 1960–1970 vs. 1990–2000[4]).

A more detailed comparison of the *NYT* subcorpora of 1900–1910 and 1990–2000 shows that the subcorpora differ in how sentence complexity is brought about. In the 1900–1910 material, though the average integration cost is somewhat higher than in 1990–2000, the individual cases involving substantial integration costs are much less frequent. In other words, reading the 1990–2000 material demands more frequent short integrations of referents, whereas the 1900–1910 material occasionally presents a giant leap of 50 words between subject and verb, but is mostly cost-free elsewhere. The *AP* corpus parallels the *NYT* in this respect.

The results remain the same when studying the cost of integration in another way, adding up the total per sentence (not per 100 words), which in effect neutralizes the impact that average sentence length may have. In effect, only the energy spent per sentence is calculated, with complete disregard to its length in words. The overall complexity remains roughly the same regardless of changes in sentence length.[5] See figures 8.1, 8.2, and 8.3 for the trend with this measure.

[3] $p<.001$

[4] The three previous decades are not statistically significant in their variation

[5] The 1990s *NYT* corpus is most complex of all the subcorpora when counted this way.

8.1.1 Types of interruption between the subject noun and verb

Relative clauses

Relative clauses between a subject head noun and verb contribute the most toward overall complexity simply because they always bring in at least one new referent, a verb, and often a new subject noun and a complement. In the *AP* and *NYT* corpora, relative clause frequency resembles the S-V complexity measures discussed in the previous section in its development over time, becoming rarer from the early part of the century until the 1960s and then increasing in the 1990s. The *AP* data ranges from 3.13 percent of all subject head nouns having relative clauses between the NH and the VP in 1900–1910 and down to 0.94 percent in the 1960s. The *NYT* ranges from 2.51 percent in the 1900s to 1.89 percent in the 1960–1970 subcorpus. The *WSP* data is again slightly different and has a low point of 2.60 percent at 1930–1940, then increasing to 2.60 in 1960–1960, and further to 3.00 percent for the 1990–2000 period.

The occurrence of object-extracted and reduced relative clauses were counted, but were not frequent enough to be statistically significant.[6]

Verbal phrases

The incidence of verbal phrases between all nominal heads and VPs follows a similar pattern in both the *NYT* and the *AP*. Especially the increase from the 1960s to the 1990s is apparent, moving from 3.41% to 4.49% for the *AP*, and from 3.51% to 4.02% for the *NYT*. Again, the *WSP* follows a different

[6]Reduced, or elliptical, relative clauses are not explicitly identified by the Machinese parser, but must be analyzed differently: a reduced RC is counted whenever a subject NH is followed by another noun (or determiner). Still, these were sporadic throughout the corpora with only one or two incidents per subcorpus, and hence not statistically significant.

pattern; there, verbal phrases increase in frequency until the 1960s, and then drop in the 1990s.

Appositives

Appositive phrase usage in the *NYT* follows the pattern set by overall complexity, becoming less frequent until the 1960s (5.86% in 1900–1910 to 4.86% in 1960–1970), then experiencing an increase in the 1990s (5.18%). The *AP* shows a more or less steady increase in appositives, moving from 1.98% of all subject-verb interpositions containing appositive phrases in 1900–1910 up to 4.80% in the 1990s. Occasionally there are several appositives in a row after the subject head noun. If every noun head of an appositive phrase is counted as a new discourse referent,[7] it certainly does increase processing requirements, but probably not as much as a relative clause, or an interposed verbal phrase.

Here again, the *WSP* follows a different trend, peaking in the 1930s and then remaining relatively steady for the 1960s and 1990s. However, the *WSP* exemplifies another use of the appositive—perhaps a more innocent contributor to complexity—which is frequently seen in the metropolitan news represented in the *WSP* corpus. Here, the appositive simply serves as a way to provide the necessary detail in a compact fashion:[8]

> It was unclear whether the car's driver, **David Daniel Heath**, **18**, was a student at the school.

This increase in appositives most likely reflects a desire to provide fine detail immediately in the news story. The appositive also lends itself to interpose

[7]This may or may not be the case, depending on the semantic distance between the noun and the appositive.

[8]Metropolitan desks at U.S. newspapers follow have the policy to, whenever possible, provide names, ages, addresses, titles, and occupations of everyday people that surface in the news. The apposition is almost always the preferred way of providing that information.

ideas that have no direct bearing on the news event, or serve as background information in the unfolding of the story. That appositives are used in this way, and that they are becoming more frequent, also leads to a loss in semantic coherence, as the the appositives often tend to split up the unity of a sentence:

> Mr. Green, a Democrat who has often criticized Mayor Rudolph W. Giuliani, a Republican, said that he would pass on the results of the survey to commissioners in the Mayor's office. (*The New York Times*)

This typical clustering of facts in a 1990s *NYT* story features two appositives: the appositive following the subject head noun introduces a relative clause, which itself carries an appositive.

8.2 New discourse referent cost

Following the simplified DLT model a "new discourse referent" is either a nominal head or a verb (not necessarily finite). Pronouns, which refer to something already existing in the discourse, were not counted.

The frequency of new discourse referents gives an indication of the information "density" of the text. The more verbs and nouns there are in relation to the total number of words, the more dense the material is. Indirectly, this also indicates the number of modifiers in relation to syntactic heads.

The trend reflects the response to the efforts of writer's guides to "be bearish on adjectives and adverbs"—all the corpora show a steady increase over time in the frequency of introducing new discourse referents. Nevertheless, in relative terms, the difference is fairly small: the largest gap is found in *AP* articles in 1900–1910 (39.13/100 words), vs. those in 1990–2000 (41.13 / 100 words)—a 4.8 percent change. However, this change in discourse density may skew the other figures somewhat. Phenomena and

cognitive load calculations have been counted in their frequency per 100 words. If 100 words in the later corpora are denser in new discourse referents, it follows that the processing cost should also be slightly higher than indicated by the cost calculations in the above sections. Reading 100 words at the end of the century would require more processing of new discourse referents, which in itself adds to the overall cost (regardless of integration cost) as the articles tend to become progressively more tightly paced.

8.3 Object noun-verb attachment

The integration of main verb and object noun only contributes a processing cost of between 0.04 to 0.12 per 100 words in the different subcorpora. Some of the cost instances were also debatable in that some nouns counted as being discourse referents between the main verb and the object noun were in fact indirect objects, but had gone untagged by the parser. This, and the relative scarcity of phenomenon, make the results not statistically significant. It should be noted, however, that when studying only the leads of 1990s stories, a concentration of object noun-verb distance is revealed.

Even though the phenomenon is rare, when it does occur, it contributes severely to the processing of a sentence, as seen in the following example—a fairly simple one in terms of subject-verb linkage, but with a number of elements interposed between the verb and the object:

> The United States Lines freighter American Surveyor will begin loading **at Exchange Place, Jersey City, this morning** 8,000 tons of food and drugs for Cuba. (*The New York Times*, 1960s)

Most "real" occurrences seem to be exceptions or reflect a lapse in construction.

	AP				NYT				WSP			
Decade	**00s**	**30s**	**60s**	**90s**	**00s**	**30s**	**60s**	**90s**	**00s**	**30s**	**60s**	**90s**
Simple NH-V constr.	28.60%	24.42%	23.15%	24.00%	26.56%	25.59%	23.08%	20.71%	26.01%	27.02%	22.79%	20.28%
Elements interposed between head noun and verb												
verbal phrases	4.41%	5.61%	3.41%	4.49%	4.67%	3.25%	3.51%	4.02%	4.18%	4.55%	5.10%	4.38%
rel cl SE	3.13%	2.80%	0.94%	2.14%	2.51%	2.87%	1.89%	4.02%	2.48%	2.09%	2.60%	3.00%
rel cl OE	0%	0%	0.23%	0.20%	0%	0%	0%	0.12%	0%	0%	0.10%	0.12%
prep phr	31.28%	24.77%	23.03%	19.19%	29.8%	26.0%	22.54%	19.79%	28.76%	26.53%	19.67%	16.13%
appositives	1.98%	3.27%	3.17%	4.80%	5.86%	5.11%	4.86%	5.18%	2.75%	3.69%	3.02%	3.11%
Integration costs												
S-V int cost (100 words)	3.78	3.34	2.94	3.12	3.76	2.95	2.58	3.20	3.11	3.13	3.10	2.78
PP int cost	5.67	5.26	5.48	6.79	5.58	6.10	6.42	6.41	6.00	5.62	5.84	6.22
S-V avg distance (in words)	2.40	2.26	2.11	2.10	2.43	2.31	2.12	2.21	2.27	2.41	2.14	2.06
New DR cost (100 words)	39.13	39.30	39.50	41.13	39.25	38.97	39.35	39.45	38.63	39.71	40.31	40.52
V-O int cost (100 words)	0.04	0.07	0.06	0.04	0.12	0.05	0.09	0.12	0.04	0.03	0.04	0.06

Table 8.1: *The main results of the study: the first row shows the percentage of subject nouns that are directly followed by the main verb. The S-V row shows the integration cost for subject head noun and verb measured in cost per 100 words on average. The following rows show the incidence of different types of phrases and clauses after a nominal head.*

Figure 8.1: *Changes in subject-verb integration cost per sentence for the AP corpus over time. The top and bottom lines at the sample points in the chart indicate the standard error of measurement, showing more variety in sentence complexity in the period 1900–1910 than in 1990–2000.*

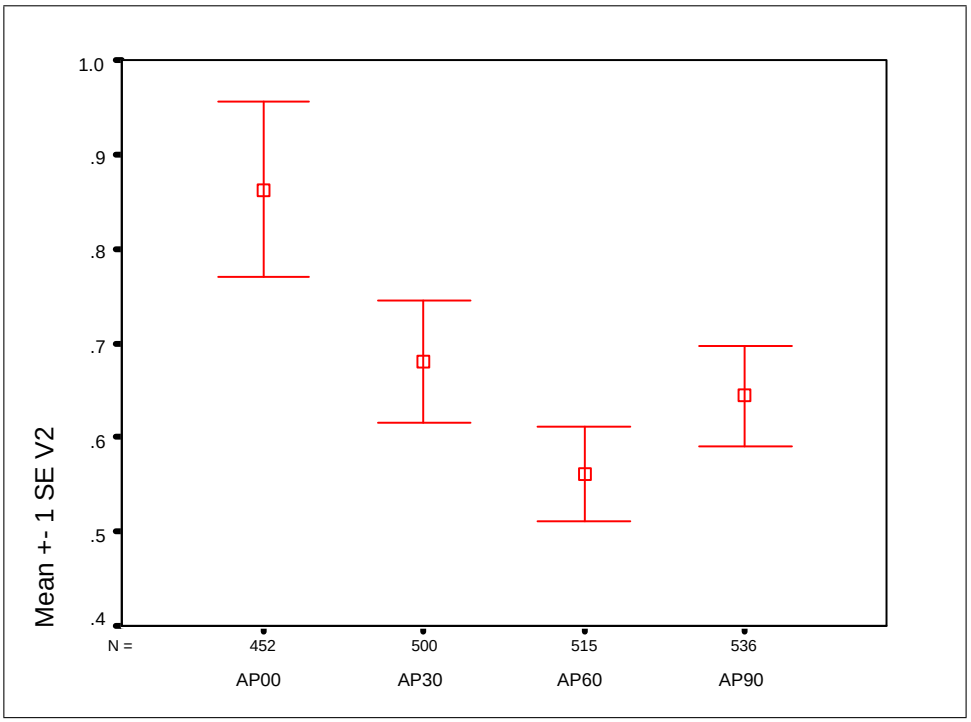

Figure 8.2: *Changes in subject-verb integration cost per sentence for the NYT corpus over time. The top and bottom lines at the sample points in the chart indicate the standard error of measurement. When complexity is counted per sentence, the NYT peaks in the 1990s.*

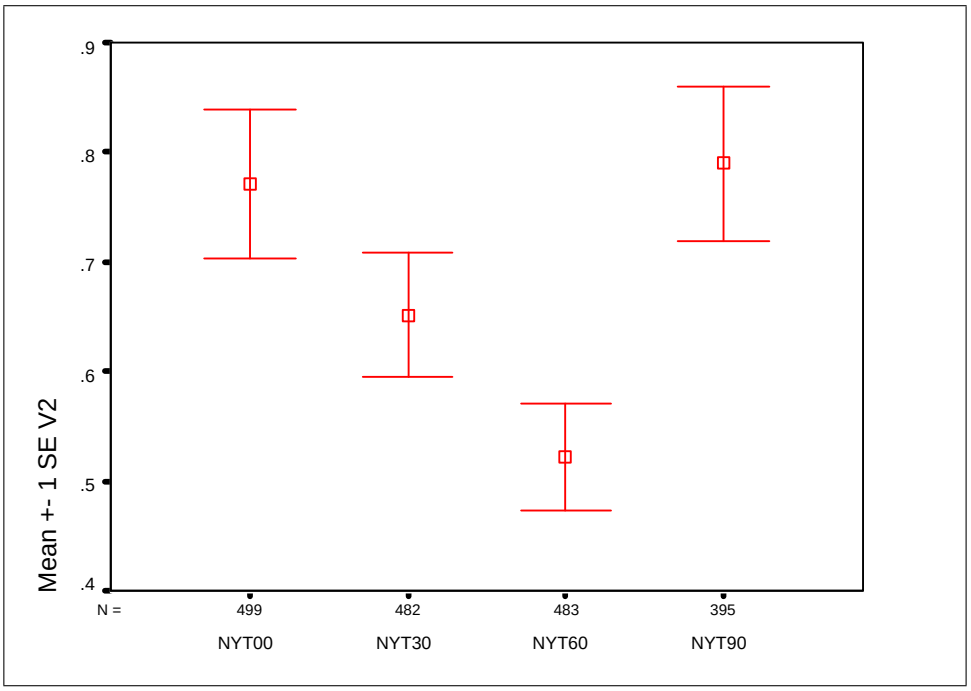

Figure 8.3: *Changes in subject-verb integration cost per sentence for the WSP corpus over time. The top and bottom lines at the sample points in the chart indicate the standard error of measurement. There is a steady decline in complexity throughout the century.*

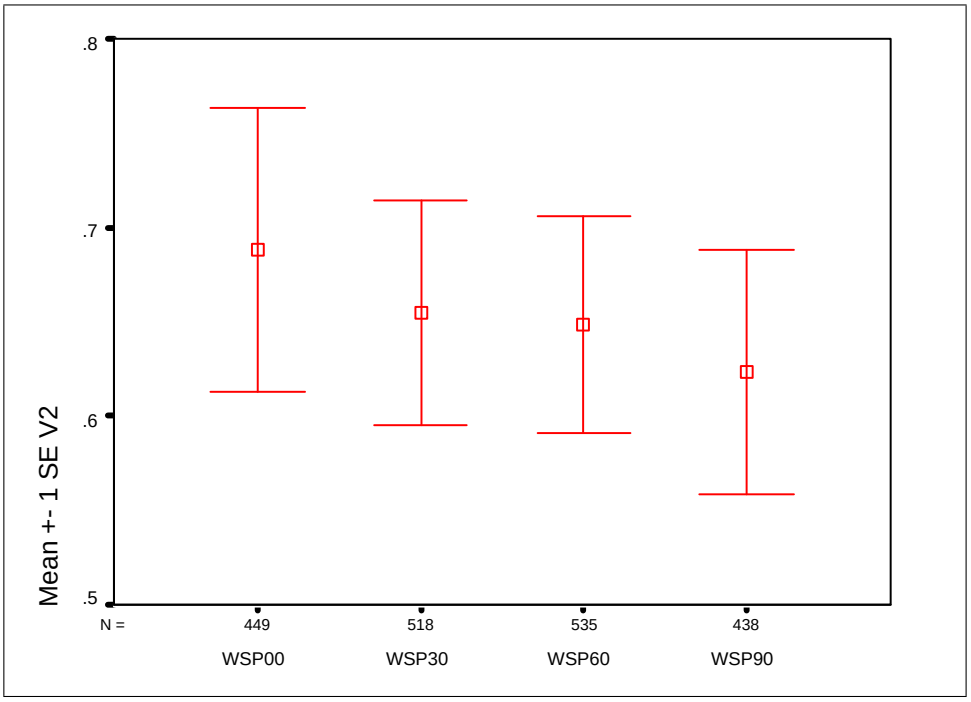

Table 8.2: *First sentences in leads in the 1990s, showing the basic complexity calculations, sentence lengths, and readability scores for the 1990s subcorpora.*

	AP	**NYT**	**WSP**
Avg sentence length	26.81	32.34	31.36
S-V int cost (100 words)	4.49	3.88	3.27
S-V avg distance (in words)	2.49	2.61	2.50
PP int cost (100 words)	6.90	6.79	7.06
Fog	18.40	20.80	21.03
Flesch	30.53	24.42	24.32
Flesch-Kincaid	15.66	17.89	17.66

8.4 The lead

Because the 1990s subcorpora of the *NYT* and the *AP* seemed to have increased in complexity, reversing the trend toward simplicity between 1900–1960, the leads (or first paragraphs) of the 1960s and 1990s stories were studied in more detail. Results show that much of the contributing complexity stems from the lead, or more precisely, from the very first sentence of a story. These first sentences had substantially higher integration costs, sentence lengths, and readability indexes than any of the remaining sections, as seen in tables 8.3 and 8.2. Leads were more "loaded" in every aspect of syntactic complexity—subject-verb attachment, PP attachment, and object-verb attachment. Here, the *AP* was found to have more of its complex sentences concentrated in the first sentence, compared with the *NYT*, which, although more complex overall, shows more distribution of complex sentences throughout a story.

Table 8.3: *First sentences in leads in the 1960s, showing the basic complexity calculations, sentence lengths, and readability scores for the 1960s subcorpora.*

	AP	**NYT**	**WSP**
Avg sentence length	24.02	26.13	27.22
S-V int cost (100 words)	2.92	3.59	5.10
S-V avg distance (in words)	2.24	2.81	3.10
PP int cost (100 words)	6.73	5.93	6.00
Fog	17.44	19.33	18.67
Flesch	29.09	24.97	28.24
Flesch-Kincaid	15.17	16.26	16.08

8.5 Vocabulary

To check for the uniformity of the vocabulary dispersion, the rate of growth of the number of unique verbs were studied.

By virtue of the nature of news text, new nouns will be introduced practically indefinitely. They come mainly in the form of proper names—descriptions of new persons and places—and are expected to pop up frequently. New nouns will surface as long as the subject matter varies. With varying subject matter, names are not likely to be repeated from one story to the next unless the corpus is very large. Accordingly, this study has not included a measure of the "closure" of nouns. In all likelihood, doing so would probably only measure the variety of subject matter in the corpus selection, rather than the dispersion of vocabulary.

Verbs, on the other hand, are quite limited in their rate of growth. They do not, as nouns do, depend heavily on the subject matter, and the introduction of new verbs tapers off fairly quickly.

As table 8.4 indicates, the rate of unique verb growth was nearly identical in all of the subcorpora. However, the earlier corpora employed slightly longer verbs. The impact of vocabulary on complexity can be held to be

Table 8.4: *The growth of unique verbs in relation to the first 100, 500, and 1000 verbs introduced in the three corpora, measured in, 1900–1910, 1930–1940, 1960–1970, 1990–2000.*

	AP				**NYT**				**WSP**			
Decade	**00s**	**30s**	**60s**	**90s**	**00s**	**30s**	**60s**	**90s**	**00s**	**30s**	**60s**	**90s**
Verbs/100	44	47	45	52	44	51	47	58	42	44	52	35
Verbs/500	172	177	169	172	152	166	176	183	169	171	186	169
Verbs/1000	267	315	279	281	247	274	296	309	282	273	292	271

quite uniform throughout the corpora.

8.6 Sentence length

Sentence length correlates to some extent, though not completely, with the other measures of complexity. In cases where complexity measures show a clear difference between subcorpora (such as *AP* 1900–1910 and 1930–1940) the same is usually also evident with sentence lengths, in this case moving from 22.88 to 20.50.[9] However, there are also cases, such as the *NYT* between 1900–1910 and 1930–1940, where S-V complexity counts show a marked change toward simplicity (3.76 to 2.95), although sentence length has grown from 19.78 to 21.88.

8.7 Readability

Readability counts follow the basic pattern set by the changes in sentence length—reflecting that measure's weight when calculating readability. Consequently, readability does not correlate in detail with the syntactic distance-based complexity measures. See table 7.1 on page 51.

[9] see table 7.1 on page 51.

Chapter 9

Discussion

> *We are in great haste to construct a magnetic telegraph from Maine to Texas; but Maine and Texas, it may be, have nothing important to communicate ... We are eager to tunnel under the Atlantic and bring the old world some weeks nearer to the new; but perchance the first news that will leak through into the broad flapping American ear will be that Princess Adelaide has the whooping cough. –Henry David Thoreau*

Examining quantitatively the syntactic complexity of the newswriting published in major American newspapers during the twentieth century, several trends become apparent. First, the writing has become more regimented—there are fewer "oddball sentences," such as giant leaps in subject-verb integration. This reflects the growing importance of "corporate" style—evidenced by the use of writing guides—over the journalist's individual preferences. Second, although the impact of the writing guides is significant, they do not hold a monopoly on dictating the style of news text. Rather, external pressures created by changes in the society and by the introduction of new media occasionally push the style into a direction that conflicts with the main advice of the guides—to write simply.

Although it is commonly assumed, in laments about declining basic reading and writing skills, that newswriting was better in an earlier day and readers were capable of understanding more complex sentences, this assumption is not entirely confirmed by the results of this study. In many ways, the use of writing guides has in fact resulted in a more coherent, more readable style. Neither have the demands on the reader consistently decreased, as evidenced by the increased complexity of the 1990s *NYT* and *AP* style.

Unlike readability formulas, linguistic models of complexity effectively reveal the differences in the *type* of complexity between the different corpora, and it is precisely this that makes the results interesting. Although complexity is again on the increase in the 1990s, it is of a very different flavor, and the newer type seems to reflect the faster pace and more disjointed nature of the late 20th century society.

9.1 Sentence length, readability, and complexity

As the vast majority of quantitative studies on clarity of language have revolved around the average sentence length and readability scores, the results at hand offer an opportunity to compare these with the results from the cognitive-theoretical methods of evaluating sentence complexity.

It has often been held that, while it's theoretically possible to write short sentences that are complex,[1] "real life" text somehow always follows the "rule" that a shorter sentence is simpler to process. The results indicate that the trend of shortening sentences has not necessarily made them easier to process; indeed, sometimes the evidence is quite to the contrary (in the *NYT* corpus, sentence length has gone up between 1900–1910 and 1930–

[1] As exemplified in the short, but unprocessable, *This is the malt that the rat that the cat that the dog worried killed ate (Yngve, 1960).*

1940, while complexity has gone down).

Because readability measures depend largely on sentence length in determining the ease of reading, the same applies to them—readability in these corpora correlates in broad measures with sentence complexity, but not enough to warrant a substitution of readability as a measure of the actual amount of mental processing a text requires.

This bears out the criticism of readability formulas—that their power is limited in predicting linguistic complexity and distinguish complex sentences from simple ones.

Also, readability formulas would not reveal much, if anything, about the different types of complexity, such as the long tangents after subject nouns in the early 20th century material and short but frequent ones in the 1990s.

9.2 The style guide

The notable increase in complexity in the 1990s is definitely something that can't be attributed to suggestions in style guides, which strongly urge the opposite. On the other hand, the 1940s campaigning for simple writing seems to have brought results: soon after it began, sentence lengths dropped, and sentence construction became more canonical SVO structures, free from subordinate clauses, verbal phrases, and heaps of prepositional phrases. The reversal of this trend which has probably begun by the 1980s [2], is most likely caused by other, more indirect suggestions and requirements.

[2]The pilot study on only *AP* material, which included the 1970s, noted that this period was even simpler than the 1960s

9.2.1 Competing priorities

The most prominent feature in the 1990s sections is the growing trend to interpose a variety of constructions immediately after the subject head noun. Sometimes this is done just to provide compact detail, as in:

> A DC-8 cargo plane, **flying for ABX Air, a subsidiary of Airborne Express**, crashed in the mountains of southwest Virginia yesterday, and the state police said there were no apparent survivors. (*AP*)

But often, the style seems connected with a need to provide background information essential to the story. This is particularly prominent in the first sentences of articles, where the importance of the news must be motivated somehow:

> Osama bin Laden, **an Arab multimillionaire whom United States officials suspect of bankrolling a network of Islamic militants**, has left his refuge in Afghanistan and flown back to the Sudan, a United States official says. (*AP*, 1990s)

This lengthening of the lead and increasing choppiness of sentences can be attributed to a number of factors—all of which do not necessarily go against the advice of style guides. Some of the qualities a story must meet are competing and mutually exclusive.

Four requirements that most current style guides say an article must fulfill may not all be congenially met:

1. *Write simply*
2. *Provide background information and a context for the story to answer the question: why is this news?*
3. *Make stories attractive and sellable from the very beginning*—this, in effect, prevents the moving of "backgrounders" to the end of the story,

or to separate paragraphs.

4. *Report all "relevant" detail*—which are conveniently interposed between subject and verb in the form of appositives and relative clauses

Observing the trend from the 1960s to the 1990s, there seems to be a tendency in the 1960s writing to sacrifice the immediate "backgrounders" in order to make the story more plain; background information is inserted in separate paragraphs toward the end of the story, or between main points. One story from 1965 begins:

> The South African Government announced today its strictest ruling yet concerning racial segregation at sports events. (*AP*)

This eight-paragraph story brings in background information about the specifics of the ruling and the South African laws little by little. Details, such as which minister had given statements on the ruling, which sports are affected, what areas of the country are affect by the law, etc. come in later. This is in violation of making the story attractive and providing an immediate context for it. Where it gains in simplicity, it loses in immediate impact.

The same simplicity is present in the early century—though explosively bloated sentences are present every now and then. One 1900s story reads:

> The Russians who recently retired from Anju to Puk-Chen are reported to have moved northward from the latter place.
>
> Twenty members of the Peddlers' Band are reported to have taken an oath to kill all officials who favor an alliance with Japan. The Japanese Minister, on being notified of this, promptly informed the Korean Government that if it did not arrest the conspirators the Japanese officials would do so. As a result four leaders of the peddlers have just been arrested. (*AP*)

It seems obvious that anyone who does not know where *Anju* or *Puk-Chen* is, and what is going on there, is excluded from the intended readership.

Many of the opening sentences, even from the early century, are surprisingly simple. But by that virtue, they count upon the reader to supply the framework to the story.

This can be contrasted with an excerpt from the 1990s New York Times corpus:

> Foreign Ministers of the European Union, meeting in Brussels, called for the immediate end of the Serbian siege of Sarajevo and said they would take "all the means necessary, including the use of air power," to get Serbian forces to lift their 22-month encirclement of Bosnia's capital.

The above introductory sentence could hypothetically be cut down to:

> *Foreign Ministers of the European Union yesterday called for the immediate end of the Serbian siege of Sarajevo.*

However, by doing so the details would be omitted—that the ministers were meeting in Brussels, that they would take "all the means necessary"—together with the background information that Sarajevo is the capital of Bosnia and that it has been besieged by Serbian forces for 22 months. Of course they could be introduced later, but that would again violate the principle to "get the news in the first paragraph."

Some style guide writers have acknowledged this dilemma, and urge simplicity to be the better choice between two evils: "A lot of facts need to be marshaled concisely, usually in the order of decreasing news value . . . writers are often tempted to overload" (Cappon, 1982).

In effect, the reporter is dealing with a see-saw, where simplicity and immediate arousal of interest are hanging at opposite ends in the balance. The outcome of this struggle between priorities is most likely determined by the journalists' perception of how much background and filler information the reader needs to grasp the importance of the story. This is line with

the standard "inverted pyramid" style—the main news must come first. But if the main news requires additional items to put it in context—in the form of reminding the reader of 22-month sieges and pointing out which capitals belong to which countries—the writing almost inevitably drifts into the realm of complex verbal acrobatics.

Parallel to this increase in complexity toward the end of the century in the sources studied here has been the introduction of another newspaper culture, exemplified particularly by *USA Today*, that strives aggressively to maintain simplicity on top of the agenda. These papers have been said to follow a mode of "broadcast" writing—producing text that largely mimicks that of television news:[3]

> For example, America's newest and highly successful national newspaper, *USA Today*, is modeled precisely on the format of television. It is sold on the street in receptacles that look like television sets. Its stories are uncommonly short, its design leans heavily on pictures, charts and other graphics, some of them printed in various colors... Journalists of a more traditional bent have criticized it for its superficiality and theatrics, but the paper's editors remain steadfast in their disregard of typographic standards. The paper's Editor-in-Chief, John Quinn, has said: "We are not up to undertaking projects of the dimensions needed to win prizes. They don't give awards for the best investigative paragraph." (Postman, 1985, p. 111)

It is interesting to note that the complexity of the "established" news outlets in this study has gone up in roughly the same period a new kind of newspaper—whose priority it is to write simply—has surfaced in the United States.[4]

This contrast reflects two responses to the "TV effect." The traditional papers strive to motivate the news with its informational context, compen-

[3]Stone (2000) found that only *USA Today* used consistently short leads that fell within the recommendations suggested by writing guides—beating other newspapers in this respect by a wide margin.

[4]Gannett, the publisher of *USA Today*, owns 101 daily newspapers in the U.S. that have a combined daily paid circulation of 7.6 million (*http://www.gannett.com*)

sating for the lack of immediacy provided in TV news by music and visual stimulation. This often leads to some added complexity of the language as background settings and developments must be provided simultaneously with the gist of an event.

The more "modern" response is to simulate the visual language of television, cutting back on textual background and instead illustating stories with photographs and graphics, keeping the text simple and firing out immediate details in a tightly paced, easily read, peek-a-boo-style.

The news sources in this study, with the exception of the *WSP*, seem to be of the former kind. The *WSP* consciously overhauled the paper in 1984 in an effort to focus on clarity and reading ease.[5] This is reflected in the continuing trend toward simplicity in the 1990s. But the more traditional news outlets are distancing and distinguishing themselves from a visual information culture by slowly drifting toward a new analytical style of reporting the news, where background facts are extensively provided and the news is constantly contextualized. This may raise the demands on the reader's attention. But—unlike the 1872 *Phileas Fogg*, who was attentive mainly to keep up with a suspense-filled narrative flow—today's reader of "traditional" newspapers will have to keep up with a complex stream of background contextual information juxtaposed with immediate news facts. The writing guides still condemn this reporting of "too many ideas," but the writers are already writing it.

[5] *The Washington Post*, http://washpost.com/gen_info/history/timeline/index.shtml

Chapter 10

Swedish Summary

Inledning

Denna avhandling är en undersökning i syntaktisk komplexitet i artiklar i två amerikanska dagstidningar och nyhetsbyrån *The Associated Press* under perioden 1900–2000. Graden av syntaktisk komplexitet har i första hand bedömts i enlighet med några samtida teorier inom kognitionsvetenskapen. För jämförelsens skull har också meningslängder och läsbarhetsindex presenterats.

De flesta tidigare studier om dagstidningsspråk har i hög grad utnyttjat sig av läsbarhetsindex för att beskriva komplexitet, eller "läsbarhet". Resultaten visar hur de kognitiva modellerna om komplexitet stämmer överens med begreppet läsbarhetsindex.

Sedan de första stilguiderna publicerats i början av 1900-talet har den amerikanska journalistiken konsekvent yrkat på ett enkelt språk i dagstidningar. Speciellt under 1940- och 1950-talen har trycket att förenkla texten varit hårt på skribenter. Studien blickar in i de olika och ibland motstridiga råd som dagstidningarnas redaktioner försökt följa.

Material

Materialet i studien består av artiklar från två amerikanska dagstidningar—*The New York Times* och *The Washington Post*—samt artiklar från nyhetsbyrån *The Associated Press*.

Artiklarna är urvalda från fyra decennier: 1900–1910, 1930–1940, 1960–1970, och 1990–2000. Utdragen består av sammanlagt av ca. 10 000 ord från varje period och källa.

Artiklarna som samlats är alla kortare än 300 ord och bär inte namnet på skribenten. Detta för att så mycket som möjligt begränsa materialet stilistiskt. Målet har varit att studera direkta reportage där ett händelseförlopp återgivits utan personliga stilistiska utsvävningar. Det utvalda materialet har skrivits av tidningarnas eller nyhetsbyråns egna journalister. Sådana artiklar där någon annan källa än de egna journalisterna uppgivits avlägsnades vid urvalet.

Artiklarna har valts ur tre sektioner: utrikesnyheter, inrikesnyheter, och regionala nyheter. Eftersom artiklarna samlats ur på måfå valda datum, återspeglar de olika sektionernas tyngd respektive källors egna utbud på de tre artikeltyperna. *The Washington Post* har rätt begränsat med egna inrikes- och utrikesnyheter, medan *AP* i stort sätt koncentrerar sig på stora inrikeshändelser och världsnyheter. *The New York Times*-materialet består av jämn uppdelning i de tre kategorierna. De *AP*-artiklar som samlats var alla sådana som verkligen funnits i tryck i någon större dagstidning; *AP*s dagliga utbud är stort, och det är inte alla artiklar som används av de tidningar som prenumererar på nyhetsbyrån.

Beräkning av komplexitet

Den syntaktiska komplexiteten i artiklarna beräknades efter de kognitiva modeller som framförts av Gibson (2000). Gibsons *Dependency Locality Theory* (DLT) har som grundprincip att största delen av den energi som går åt till att konstruera betydelsen av en mening används för att förena syntaktiska element. Ju längre isär de syntaktiska elementen står, desto svårare blir uppgiften att parsa en mening. Oftast är den kognitivt sett tyngsta uppgiften att koppla ihop huvudordet i en sats till verbet.

Gibsons teori har i studien simplifierats enligt ett förslag i Gibson (2000). För att beräkna komplexitet studerar man då endast antalet nya diskursreferenter som uppkommer mellan två syntaktiska element. I meningen

> The reporter who sent the photographer to the editor hoped for a good story

skall huvudordet **The reporter** kopplas ihop till verbet **hoped**. Detta förbrukar tre s.k. integrationsenheter, eftersom det emellan de två orden placerats tre nya diskursreferenter: **sent**, **the photographer**, och **the editor**.[1]

Allt material parsades med Connexor's *Machinese Syntax* programvara. Programmet producerar morfosyntaktiska taggar ifrån vilka man kan räkna ut hur de syntaktiska elementen hör samman. Studien koncentrerar sig på distansen mellan tre olika typer av sammanfogning: huvudord-verb, verb-objekt, samt prepositionsfraser.

Av största vikt är dock avståndet mellan huvudord och verb. Typen av satser och konstruktioner som faller emellan de två har studerats i ytterligare detalj.

[1] Alla verb och substantiv räknas som nya diskursreferenter. Däremot räknas inte pronomen, då de redan hänvisar till något som lagts fram tidigare.

Resultat

The New York Times och *AP* är rätt lika i sin utformning under de fyra tidsperioderna—*NYT* skriver i stort sett något mer komplext, men utvecklingen följer samma mönster hos de två. Stilen hos *AP* och *NYT* är mest invecklad under 1900–1910, och blir sedan märkbart enklare ända t.o.m. 1960–1970. I 1990–2000 är satskonstruktionerna dock igen något mer komplexa—i *NYT* kanske mer än någonsin tidigare, beroende på om man räknar den sammanlagda integrationskostnaden per ord eller per mening.

The Washington Post avviker något från de två andra källorna i att trenden genomgående pekar på sänkning i komplexiteten.

I alla tre källorna märks en klar tendens att i senare tider undvika en kanonisk SVO meningsstruktur. Antalet meningar där verbet inte direkt efterföljer huvudordet har ökat ständigt. Detta betyder att bisatser, verbfraser, och prepositionsfraser oftare sjuts in mellan huvudord och verb i de senare korpusarna, men att de å andra sidan blivit allt kortare, och därmed inte bidrar lika mycket till komplexiteten.

Läsbarhet, meningslängd, och komplexitet

Läsbarheten i alla korpusarna studerades också och jämfördes med de övriga observationerna. I stort sett följer läsbarhetsindex och meningslängd också de kognitiv-baserade komplexitetsmätningarna. Men undantag förekommer. Till exempel *NYT* korpusen har enligt läsbarhetsindex blivit mer svårläst mellan 1900–1910 och 1930–1940.[2] Enligt komplexitetsobservationerna har dock meningarna under denna tid blivit märkbart mindre komplexa. Skillnaden beror antagligen på att meningarna har blivit längre och att läsbarhetsindex stöder sig främst på meningslängd i beräkning av komplex-

[2]Detta stämmer också överens med tidigare forskning, se kapitel 6.

itet.

Sammanfattning

Den kvantitativa undersökningen av syntaktisk komplexitet i amerikanska dagstidningar påvisar några tydliga trender. Texten har under århundradets lopp blivit mer formelföljande—det finns färre meningar som avviker från mängden i slutet av 1900-talet. Trots detta följer inte stilen helt instruktionerna som riktas åt journalister: 1990-talet visar ett uppsving i komplexiteten efter en tidigare stadig trend mot förenkling mellan 1900 och 1970. Detta går emot de härskande råden i branchen—att till varje pris skriva klart och enkelt. Å andra sidan kan man konstatera att den starka kampanj för enkelt språk i dagstidningar som påbörjades på 1940-talet nog hade en märkbar betydelse. Mellan 1930–1940 och 1960–1970 har en synbar förenkling skett.

I motsats till läsbarhetsindex avslöjar forskning i komplexitet med hjälp av språkvetenskapliga modeller (som DLT) vilka fenomen som mest bidrar till skillnader mellan korpus och vilken typ av vanliga konstruktioner som gör meningarna i nyhetstext invecklade. Trots att den syntaktiska komplexiteten i artiklar har ökat mot 1990-talet, visar det sig att de meningar som kan anses vara invecklade är i stort sett av en annan typ än under början av århundradet. *AP* och *The New York Times* korpusarna 1900–1910 visar mycket större variation i stil än de senare samlingarna. Oftast är de tidigare meningarna rätt så korta, med huvudord som ligger nära verbet. Men då och då förekommer mycket komplicerade meningar som bidrar till att genomsnittskomplexiteten ligger högre än under resten av århundradet. På 1990-talet, där komplexiteten igen ökat gentemot 1960-talet, finner man inte längre ett lika brett register: relativa bisatser, appositioner, och verbfraser skjuts regelbundet in mellan huvudordet och verbet, men dessa är ofta rätt

korta konstruktioner och genomsnittliga komplexiteten når inte helt upp till 1900–1910 korpusens standard.

Vid närmare undersökning visar det sig att de meningar som starkast bidrar till komplexitet ofta är sådana där skribenten vill presentera bakgrundsstoff för läsaren samtidigt som den centrala nyheten berättas. Ett exempel från 1990-talet lyder:

> Foreign Ministers of the European Union, meeting in Brussels, called for the immediate end of the Serbian siege of Sarajevo and said they would take "all the means necessary, including the use of air power," to get Serbian forces to lift their 22-month encirclement of Bosnia's capital.

Den första meningen, som enligt den traditionella nyhetsmodellen skall besvara alla centrala frågor om artikeln, utnyttjas här till fullo. Läsaren får indirekt veta att Sarajevo är huvudstad i Bosnien, att belägringen pågått i 22 månader, att det är serberna som är belägrare, och att EU:s utrikesministrar i allmänhet håller möte i Bryssel.

Däremot finner man i 1900–1910 korpusen artiklar som följande:

> The Russians who recently retired from Anju to Puk-Chen are reported to have moved northward from the latter place.
>
> Twenty members of the Peddlers' Band are reported to have taken an oath to kill all officials who favor an alliance with Japan. The Japanese Minister, on being notified of this, promptly informed the Korean Government that if it did not arrest the conspirators the Japanese officials would do so. As a result four leaders of the peddlers have just been arrested.

Här inskjutes inga relevanta fakta parallellt med själva nyheten utan läsaren har tydligen själv ansvaret att hitta ett sammanhang för händelserna i *Anju* och *Puk-Chen*, och resten av artikeln.

Trenden att erbjuda mer bakgrundsinformation tillsammans med själva nyheten är märkbar just i *AP* och *NYT*. Det står nära till hands att konstatera att skribenterna mot slutet av århundradet undviker det ständigt

yttrade rådet att skriva enkelt, med korta meningar och utan bisatser efter huvudord. Dock finner man att stilguiderna inte är fullt så entydiga med rådet; det finns en mängd andra krav på artiklar som delvis står i strid med strävan till enkelhet. Framför allt måste skribenten också ta vara på att placera nyheten i ett sammanhang och att genast erbjuda de detaljer som gör nyheten intressant. Det rör sig då om att finna en balans mellan de olika krav som ställs på skribenten vid dagstidningarna och nyhetsbyråerna. Å ena sida skall texten vara enkel, meningarna korta, helst utan avbrott mellan huvudord och verb. Men läsaren skall samtidigt också presenteras en koherent bakgrund till händelserna som ligger vid kärnan av en rapport. Ytterligare skall också första paragrafen i en artikel presentera alla de relevanta detaljerna, och samtidigt motivera för läsaren varför artikeln i fråga är en nyhet. Dessa olika krav kan kanske inte tillfredställas på ett flytande sätt.

Texterna på 1960-talet uppvisar ofta en tendens att utelämna både kontext och detaljer. Man finner dem ofta i egna paragrafer långt efter huvudsaken har kommit fram. Då behövs sällan heller komplicerade meningar där alla subjekt vidareförklaras parallellt med framförandet av händelserna. I och med detta uppoffras ju dock några av rekommendationerna för nyhetstext.

På 1990-talet vävs däremot bakgrunden ofta in samtidigt med händelseförloppet. Attribut och bisatser faller då mellan huvudord och verb och gör meningarna komplexa. Detta kan ses som en alternativ lösning på dilemmat om de motstridiga kraven på nyhetstext.

Mellan 1980–2000 har det också skett en uppdelning av dagstidningskulturen i USA. Tidningar som *USA Today* har förvärvat en bred läsarbas med att aggressivt hålla sig till ett förenklat språk och bifoga rikliga illustratio-

ner till artiklar. Tidigare studier på 1990-talet har visat att just denna nya tidningstyp är i stort sett den enda som håller sig till de riktlinjer om enkla meningar som branchen själv rekommenderar. Att komplexiteten samtidigt börjat tillta i de mer traditionella och "allvarliga" nyhetskällorna *AP* och *The New York Times* har förstärkt denna splittring av nyhetskulturen.

Appendix A

Machinese Syntax Tags

A.1 English syntactic relations

Tag	**Explanation**	**Example**
main	main nucleus of the sentence; usu. main verb	
qtag	tag question	It is cold, isn't it?
v-ch	verb chain: auxiliaries + main verb	
pm	preposed marker. marker of a sub. cl.	
pcomp	prepositional complement	They are in that red *car*.
phr	verb particle	She looked *up* the word.
subj	subject	*John* is in the kitchen.
agt	agent	The agent by-phrase in passive sentences.
obj	object	
comp	subject complement	John is *foolish*.
dat	indirect object	John gave *him* an apple.
oc	object complement	John called him a *fool*.
copred	copredicative	John regards him *as* foolish.
com	comitative	Drinking *with* you is nice.
voc	vocative	*John*, come here!
ins	instrument	He sliced the salami *with* the knife.
tmp	time	

dur	duration	
frq	frequency	
qua	quantity	
man	manner	
loc	location	
sou	source	
goa	goal	
pth	path	He travelled from Tokyo to Beijing.
cnt	contingency	
cnd	condition	
meta	clause adverbial	*So far*, the OECD has refused.
cla	clause initial adverbial	*Under* Clinton, the economy is more regulated.
ha	heuristic prepositional phrase attachment	The beam will escape *through* the mirror.
qn	quantifier	
det	determiner	
neg	negator	
attr	attributive nominal	
mod	other postmodifier	
ad	attributive adverbial	
cc	coordination	Jack *and* Jill bought some pins, nails *and* needles.

A.2 English surface syntactic tags

Tag	Explanation	Example
@+FAUXV	finite auxiliary predicator	
@-FAUXV	nonfinite auxiliary predicator	
@+FMAINV	finite main predicator	
@-FMAINV	nonfinite main predicator	
@SUBJ	subject	
@F-SUBJ	formal subject	
@OBJ	object	
@I-OBJ	indirect object	John gave *him* an apple.
@PCOMPL-S	subject complement	
@PCOMPL-O	object complement	
@ADVL	adverbial	
@O-ADVL	object adverbial	She let him walk *the streets*.
@APP	apposition	
@NH	stray noun phrase	
@VOC	vocative	*John*, come here!
@A>	premodifier of a nominal	
@DN>	determiner	
@QN>	premodifying quantifier	
@AD-A>	intensifier	
@<NOM-OF	postmodifying prepositional phrase beginning with of	
@<AD-A	postmodifying intensifier	
@<NOM	postmodifier of a nominal	
@INFMARK>	infinitive marker to	
@<P-FMAINV	nonfinite clause as preposition complement	
@<P	other preposition complement	
@CC	coordinating conjunction	
@CS	subordinating conjunction	
@DUMMY	unspecified	

A.3 English functional tags

Tag	**Explanation**	**Example**
&NH	nominal head	
&N<	postmodifier of a nominal	
&>	premodifying adverb	
&AH	adverbial head	
&A<	postmodifying adverb	
&AUX	auxiliary verb or particle	
&VP	main verb in a passive verb chain	
&VA	main verb in an active verb chain	
&>CC	introducer of coordination	
&CC	coordinating conjunction	
&CS	subordinating conjunction	

A.4 English morphological tags

Part of speech	**Subfeature**	**Explanation**
N		noun
—case	NOM	nominative
	GEN	genitive
—number	SG	singular
	PL	plural
ABBR		abbreviation
—case and number like in nouns		
A		adjective
—comparison	ABS	absolutive
	CMP	comparative
	SUP	superlative
NUM		numeral
	CARD	cardinal
	ORD	ordinal
—number	SG	fraction, singular
	PL	fraction, plular
PRON		pronoun
—case and related features	NOM	nominative
	GEN	genitive
	ACC	accusative
	INDEP	the independent genitive form (eg. theirs)
—number	SG	singular
	SG1	singular, first person
	SG3	singular, third person
	PL	plural
	PL1	plural, first person
	PL3	plural, third person
—comparison	ABS	absolutive
	CMP	comparative
	SUP	superlative

—other pronoun subfeatures	PERS	personal
	DEM	demonstrative
	RECIPR	reciprocal
	WH	rel or interr pron beginning with the wh or how
	<Interr>	interrogative
	<Refl>	reflexive
	<Rel>	relative
DET		determiner
—case	GEN	genitive
—number	SG	singular
	PL	plural
—comparison	ABS	asbolutive
	CMP	comparative
	SUP	superlative
—other subfeatures of determiners	DEM	demonstrative
	WH	det beginning with wh or how
ADV		adverb
—comparison	ABS	absolutive
	CMP	comparative
	SUP	superlative
—other subfeatures for adverbs	<Ex>	existential *there*
	WH	adverb beginning wh or how
ING		present participle
EN		past participle
V		verb; used only for finite verbs and infinitives
	AUXMOD	modal auxiliary
	INF	infinitive
	IMP	imperative
	SUBJUNCTIVE	subjunctive
—tense	PRES	present tense
	PAST	past tense
—person	SG1	singular, first person

	SG3	singular, third person
—other subfeatures for verbs	<N+>	N-V combination (e.g. India's)
INTERJ		interjection
CC		coordinating conjunction
CS		subordinating conjunction
PREP		preposition
NEG-PART		the negative particle
INFMARK>		infinitive marker
<?>		mark for unknown word

Appendix B

A Glossary of News Terminology

Angle	The particular point of emphasis in a story, often given in the lead.
Byline	The reporter's signature that heads a story.
Cablese	Short abbreviated language typical of telegraphic transmission.
Color	Words and sentences that are concrete and sensory
Copy	Applies to all written material.
Cover	To report an event
Credit line	Place in a story where attribution to outside source is given.
Dateline	The line giving the origin of the story.
Desk	A division within a newspaper (metropolitan desk, foreign desk, etc.)
File	To transmit a story on the wire.
Lead	The first paragraph of a story—often a single sentence.
Pickup	A story where the facts are attributed to another news source.
Slug	A short mark that identifies a story (e.g. Vietnam–Casualties).
Spot news	Immediate news, often of an unexpected event.
Straight news	An unembellished recital of factual material.
Style sheet	An organization's compilation of in-house style rules.
Style	The practices of punctuation and spelling of a particular news outlet.
Wire	The telegraph wire where news agency stories arrive.

Bibliography

Anderson, R. C. and Davidson, A. (1988). Conceptual and empirical bases of readability formulas. In Davidson and Green (1988), pages 26–54.

Asprey, M. (2003). *Plain Language for Lawyers*. The Federation Press, Sydney, Australia.

Baker, E. L., Atwood, N. K., and Duffy, T. M. (1988). Cognitive approaches to assessing the readability of text. In Davidson and Green (1988), pages 55–84.

Bates, S. (1989). *If No News, Send Rumors: Anecdotes of American Journalism*. Henry Holt and Company, New York.

Bruce, B. and Rubin, A. (1988). Readability formulas: Matching tool and task. In Davidson and Green (1988), pages 5–21.

Burgoon, J., Burgoon, M., and Wilkinson, M. (1981). Writing style as predictor of newspaper readership, satisfaction and image. *Journalism Quarterly*, 58(2):225–231.

Bush, C. R. (1954). *The Art of News Communication*. Appleton-Century-Crofts, New York.

Campbell, L. R. and Wolseley, R. E. (1961). *How to Report and Write the News*. Prentice-Hall, Englewood Cliffs, N.J.

Cappon, J. (1990a). More on sentence length. *The [AP] Insider*, 2(3).

Cappon, J. (1990b). On simple sentences and complex clutter. *The [AP] Insider*, 2(5).

Cappon, J. (1991). The l-word. *The [AP] Insider*, 3(4).

Cappon, R. J. (1982). *The Word: An Associated Press Guide to Good News Writing*. The Associated Press, New York.

Chall, J. S. (1988). The beginning years. In Zakaluk and Jay (1988), pages 2–13.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press, Cambridge, Massachusetts.

Church, K. and Patil, R. (1982). Coping with syntactic ambiguity or how to put the block in the box on the table. *American Journal of Computational Linguistics*, 8(3–4):139–149.

Danielson, W. A. and Bryan, S. D. (1964). Readability of wire stories in eight news categories. *Journalism Quarterly*, 41:105–106.

Davidson, A. and Green, G., editors (1988). *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Doelling, O. C., editor (1998). *The Associated Press Handbook for International Correspondents*. The Associated Press.

Fenby, J. (1986). *The International News Services*. Schocken Books, New York.

Flesch, R. F. (1943). Estimating the comprehension difficulty of magazine articles. *Journal of General Psychology*, 28:63–80.

Foulger, D. (1978). A simplified flesch formula. *Journalism Quarterly*, 55:167,202.

Fowler, G. L. (1978). The comparative readability of newspapers and novels. *Journalism Quarterly*, 55:589–592.

Frank, R. (1992). *Syntactic locality and tree-adjoining grammar: Grammatical, acquisition, and processing perspectives.* PhD thesis, University of Pennsylvania, Philadelphia.

Fry, E. B. (1988). Writeability: The principles of writing for increased comprehension. In Zakaluk and Jay (1988), pages 77–95.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68:1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain. Papers from the first Mind Articulation Project Symposium*, pages 95–126.

Gibson, E., Timothy, D., Grodner, D., Watson, D., and Ko, K. (2004). Reading relative clauses in english. *Cognitive Linguistics (in press).*

Gramling, O. (1968). *AP—The Story of News.* University Microfilms, Ann Arbor, Michigan.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency.* Cambridge University Press, Cambridge.

Hoskins, R. L. (1973). A readability study of ap and upi wire copy. *Journalism Quarterly*, 50:360–363.

Hunt, K. W. (1964). *Differences in grammatical structures written at three grade levels, the structures to be analyzed by transformational methods.* U.S. Department of Health, Education, and Welfare, Tallahassee, Florida.

Hyde, G. M. (1912). *Newspaper Reporting and Correspondence.* D. Appleton and Co., New York.

Hyde, G. M. (1952). *Newspaper Reporting.* Prentice-Hall, Englewood Cliffs, N.J.

Jones, J. P. (1949). *The Modern Reporter's Handbook.* Rinehart and Co., New York.

Jung, J. and Jo, S. (2001). A comparative analysis of on-line versus print media: Readability and content differentiation of business news. *2001 AEJMC Convention.*

Kemper, S. (1988). Inferential complexity and the readability of texts. In Davidson and Green (1988), pages 141–166.

Kintsch, W. A. and Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In Nilsson, L.-G., editor, *Perspectives on memory research*, pages 329–265, Hillsdale, New Jersey. Erlbaum.

Klare, G. R. (1963). *The Measurement of Readability.* Iowa State University Press, Ames, Iowa.

Lively, B. and Pressey, S. L. (1923). A method for measuring the 'vocabulary burden' of textbooks. *Educational Administration and Supervision*, 9:389–398.

Manning, C. D. and Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. MIT Press.

Miller, G. A. and Chomsky, N. (1963). Finitary models of language users. In Luce, R. D., Bush, R. R., and Galanter, E., editors, *Handbook of Mathematical Psychology, Vol. II*, pages 419–491. John Wiley, New York.

Postman, N. (1985). *Amusing ourselves to Death: Public Discourse in the Age of Show Business*. Penguin.

Randall, J. H. (1988). Of butchers and bakers and candlestick-makers: The problem of morphology in understanding words. In Davidson and Green (1988), pages 223–246.

Razik, T. A. (1969). A study of american newspaper readability. *Journal of Communication*, 19:317–324.

Schwarzlose, R. A. (1979). *The American Wire Services*. Arno Press, New York.

Schwarzlose, R. A. (2002). Cooperative news gathering. In Sloan and Parcell (2002), pages 153–162.

Sloan, D. W. and Parcell, L. M., editors (2002). *American Journalism—History, Principles, Practices*. McFarland, Jefferson, North Carolina.

Stone, G. (2000). Lead length and voice in u.s. newspapers. *Web Journal of Mass Communication Research*, 3.

Tapanainen, P. (1999). *Parsing in two frameworks: finite-state and functional dependency grammar*. Academic Dissertation. University of Helsinki, Language Technology, Department of General Linguistics, Faculty of Arts., Helsinki.

Tapanainen, P. and Järvinen, T. (1997). A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71, Washington, D.C. ACL.

Vos, T. P. (2002). News writing structure and style. In Sloan and Parcell (2002), pages 296–305.

Westley, B. (1953). *News Editing.* Houghton Mifflin, Cambridge, Massachusetts.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.

Zakaluk, B. L. and Jay, S. S., editors (1988). *Readability: Its Past, Present & Future.* International Reading Association, Newark, Delaware.