Processing Internet-derived Text—Creating a Corpus of Usenet Messages

Sebastian Hoffmann
Department of Linguistics and English Language, Bowland College, Lancaster University

Abstract

In recent years, linguists have become increasingly interested in the language of the Internet—both as an object of investigation as well as a source of authentic data to complement traditional electronic corpora. However, Internet-derived data is typically very messy data and a conversion process is often required in order to enable researchers to carry out a reliable quantitative investigation of the patterns observed with the help of standard corpus tools. In this article, I discuss the technical and methodological aspects involved in creating a large corpus of asynchronous computer-mediated communication by downloading and post-processing hundreds of thousands messages posted in twelve Usenet newsgroups. After describing how messages can be arranged into hierarchically structured discussion threads, I focus at some length on the strategies that are required to correctly assign authorship to the different textual elements in individual messages. My algorithms have a success rate of well over 90% for most newsgroups and the resulting corpus can thus serve as a suitable basis for an investigation into the interactive strategies employed in this particular type of written communication.

Correspondence:
Sebastian Hoffmann,
Department of Linguistics and English Language,
Bowland College,
Lancaster University,
Lancaster LA1 4YT, UK.
Email:
s.hoffmann@lancaster.ac.uk

1 Introduction

Given the enormous growth of the World Wide Web over the past fifteen years, it is no surprise that linguists increasingly turn to the Internet as a source of authentic language data in order to complement the more traditional electronic corpora available today. For this purpose, they usually pursue one of several possible options.[1] The most readily accessible of these is to consider the whole of the World Wide Web as a single corpus that can be accessed by way of commercial search engines such as Google or Altavista. However, while a cautious interpretation of the results of such searches may indeed lead to linguistically interesting findings (see for example Mair, 2003, 2007; Rohdenburg, 2007), a number of major drawbacks exist that seriously restrict the value of such Web-based searches for general use. Thus, whereas traditional corpora are carefully compiled in order to be used as a representative sample of the language, the composition and size of the World Wide Web is largely unknown and is constantly changing. As a consequence, search results cannot be reliably replicated and normalized frequency counts (e.g. for comparisons between different text categories) are not available. In addition, the search algorithms implemented by commercial search engines are often insufficient for the requirements of linguistic retrieval, and the relevance-based ranking of the query result renders a reliable interpretation of the findings difficult.[2]

Literary and Linguistic Computing, Vol. 22, No. 2, 2007.
doi:10.1093/llc/fqm002 Advance Access Published on 28 February 2007

Another option for the linguistic exploitation of the Internet is to restrict the object of investigation to a clearly defined subsection of the World Wide Web. For example, many newspapers offer free online access to selected extracts of their print versions and extensive Web archives are often available that can be searched with the help of basic search facilities.[3] In contrast to general searches of the whole Web, this approach offers the advantage of working with a more clearly defined set of data (e.g. British newspaper language). However, many of the other drawbacks such as the lack of normalized frequency counts and the necessary reliance on non-linguistic search engines remain.[4]

The third option for making use of online resources is to create a local copy of the data in question. Once the relevant Web pages have been downloaded, they can be post-processed to suit the needs of scholars and searched with the corpus tools of their choice. In addition, since it is possible to determine the exact amount of data that is being searched, normalized frequency counts can easily be calculated. Several large-scale projects are currently underway which aim at making sizeable portions of the World Wide Web accessible via a linguistic search engine (cf. Kilgarriff, 2003; papers in Baroni and Bernardini, 2006). However, while access to such large collections of Web pages will eliminate some of the more serious drawbacks of Internet-based investigations, scholars may often need access to a more homogeneous and ordered set of data. For this purpose, much smaller and tailor-made Web-derived corpora will often allow researchers to greatly expand the range of available data without them having to compromise on the application of standard corpus linguistic methodology.

The present article will focus on the methodological and technical aspects involved in creating such a specialized Internet-derived corpus. More specifically, I will concentrate on Usenet forums as a valuable source of interactive language use. Usenet consists of thousands of online discussion groups in which users exchange messages about a very large number of different topics. While such a corpus of course cannot be treated on a par with the spoken corpora currently available, the strategies employed in such written exchanges may nevertheless provide interesting additional insights into the nature of verbal interaction.

The procedure described in this article requires some skills in the programming language Perl. With its powerful regular expression engine, Perl lends itself very well to the manipulation of large amounts of text.[5] This requirement may deter some less computer-savvy readers from attempting to build their own Internet-derived corpora. However, it is one of the aims of this article to demonstrate that even Perl scripts of a low level of sophistication—often referred to as 'dirty hacks'—can lead to the creation of very useful data resources. It is perhaps worth pointing out that I am a linguist who has never had any formal education in programming. Instead, my programming skills are self-taught and were developed according to the needs established by the linguistic tasks at hand. Readers are therefore explicitly encouraged to consider acquiring some basic programming skills on their own.[6]

2 Usenet

Usenet consists of a very large number of thematically organized, public message boards (or 'newsgroups') which are in principle accessible to anybody with a computer that is connected to the Internet. Established in 1979, Usenet pre-dates the World Wide Web with its hypertext-based graphical user interface by more than a decade.[7] Newsgroups are organized hierarchically under a number of top-level domains (e.g. 'soc' for 'social issues and personal interaction'; 'rec' for 'hobbies, arts, and recreational activities') and their names are indicative of the topics that are discussed. Thus, the newsgroup soc.culture.malaysia offers participants a forum to converse about matters relating to Malaysian life and culture whereas soc.culture.thailand and soc.culture.singapore will offer similar topic areas concerning Thailand and Singapore, respectively. In recent years, the popularity of Usenet has somewhat declined—perhaps as a result of the increased use of e-mail based discussion group services and virtual diaries (blogs).

```
Newsgroups: alt.usage.english
From: "Don Phillipson" <d.phillipson@ttrryytteell.com>
Date: Sun, 3 Jul 2005 17:55:05 -0400
Subject: Re: contact with someone or contact someone?

"walker" <abc@xyz.com> wrote in message
news:da9hpk$jcl$1@domitilla.aioe.org...

> I googled and found both. Which one is used more widely?

Googling for one or two words helps little because
Google cannot differentiate the verb contact from
the noun contact, e.g.

1.  I am in contact (noun) with XYZ.
2.  I want to contact (verb) XYZ.

Both forms are currently used.
```

Fig. 1 A message posted on the newsgroup alt.usage.english

In contrast to Internet Relay Chat (IRC; cf. Werry, 1996), which requires participants to be online simultaneously, Usenet forums belong to the category of asynchronous computer-mediated communication (CMC). Thus, messages do not require immediate attention and are often replied to hours or even days after they were originally posted.[8] However, Usenet discussions are nevertheless clearly interactive in nature. This impression is supported by the fact that participants have the option of quoting passages from previous posts as part of their replies. This greatly facilitates the establishment of topical coherence as the relevance of a particular statement can be clearly indicated by way of a sequential ordering of old and new elements—even though a considerable lapse of time may exist between the writing of the original message and its reply.[9] Usenet discussions are thus a hybrid form of communication in the sense that they combine features of face-to-face talk with those of written texts.

Figure 1 displays a typical example of a Usenet message posted on the newsgroup alt.usage.english. Here, the author replies to a question about the correct use of the expressions *contact with* and *contact someone* and has chosen to integrate the relevant part of the original message (i.e. 'I googled and found both. Which one is used more widely?') into his own contribution. Instead of quotation marks, an angle bracket placed at the beginning of the line is used to distinguish the quoted passage from the new elements of text.

The language of Usenet has received comparatively little scholarly attention. A number of studies exist that make use of fairly small sets of data to present a qualitative analysis of some newsgroup-specific aspects of language use. For example, while some investigations focus on the strategies that are conditioned by the asynchronous and public nature of this particular type of CMC (e.g. Lewin and Donner, 2002; Marcoccia, 2004), others study the formal properties of the language found in a number of specific newsgroups and describe how it differs from standard grammar (e.g. Gheno, 2003, for Italian). Further areas of interest are the representation of gender differences (Buchanan, 2000) as well as the argumentative techniques used in agreements and disagreements (Baym, 1996). A much bigger corpus of 11,176 messages is used in Galegher et al.'s (1998) analysis of the rhetorical strategies employed in exchanges found on three electronic support groups. However, it is only within the field of computational linguistics that Usenet data has been used on a large scale. Thus, scholars have successfully utilized a corpus of Usenet discussion groups as a proxy for spoken language for the purpose of automatically building a system for identifying colloquial vocabulary in Chinese (Cheung and Fung, 2004). Also, a 300-million word

corpus of newsgroup data has been used to develop strategies for detecting and resolving semantic ambiguities (Burgess, 2001).

The interactive nature of newsgroup discussions clearly deserves to be investigated in greater detail. In particular, quantitative studies involving a large number of postings from a range of different newsgroups may contribute to a better understanding of the strategies employed in this particular type of CMC. Such studies will also enable researchers to offer a more reliable assessment of the parallels between face-to-face conversations and the type of written conversations found on Usenet. However, for such an undertaking to be successful, the data has to exist in a format that makes it possible to distinguish reliably between contributions from different users. While human readers usually face little or no difficulties in assigning different stretches of text to different writers, an automatic analysis of the data unfortunately presents more complex challenges. In fact, as soon as several authors take part in the same discussion, Usenet messages tend to turn into very messy data. Considerable effort is then required to isolate the individual 'turns' contributed by different people. In the following sections, I will illustrate the steps that are necessary to create a large corpus of Usenet messages which will not only be fully searchable with the help of standard corpus tools but which will also allow researchers to reliably reconstruct the exact (temporal) development of the electronically stored 'conversations'.

3 Downloading the Data

The first step in compiling a corpus of Usenet data consists of downloading all available messages from a selection of newsgroups to a local computer. As is the case for any type of corpus compilation, this selection process is a crucial step which greatly influences the range and quality of the linguistic results that can be obtained on the basis of the final corpus. Linguists who are interested in a representative coverage of topics discussed on Usenet may wish to include a large number of different newsgroups from all existing top-level domains.

Table 1 Newsgroups downloaded

Name	*N* messages	*N* words
alt.alien.research	58,189	13,978,358
alt.coffee	68,056	12,076,503
alt.fan.noam-chomsky	62,703	25,683,845
alt.games.warcraft	23,849	3,896,371
alt.music.oasis	21,883	1,934,548
news.software.nntp	5,033	1,177,785
news.software.readers	31,874	4,790,531
rec.audio.tech	24,645	5,247,679
rec.gambling.sports	34,848	4,955,744
rec.music.classical.recordings	147,866	28,188,853
rec.photo.digital	279,846	46,358,859
rec.sport.swimming	15,980	3,360,974
Total	773,772	151,650,050

Others may want to concentrate on a particular group of topics or a clearly delineated set of participants and therefore select only a small number of different groups.[10] Since the focus of this study is on the technical issues involved in the creation of a Usenet corpus, the composition of the final set of data is of course less important here. For demonstration purposes, I have selected a total of twelve fairly high-volume groups which belong to three different top-level domains. The complete list is shown in Table 1, which also displays the number of different messages per newsgroup and the total number of words downloaded.

There are several options for downloading Usenet messages to a local computer. Standard newsreaders—which are available for all operating systems—typically allow users to export messages to the hard disk. This is a viable option if the contents of only a small number of newsgroups need to be downloaded. Researchers who are looking for a broad coverage of newsgroups may consider setting up a news server on their own. This typically requires Unix system administration skills as well as access to a commercial newsfeed. The complete contents of all selected newsgroups are then 'sucked' onto the local hard disk. This process can result in the transfer of enormous amounts of data and thus requires adequate hardware and network bandwidth. For my own purposes, a short Perl script that essentially functions as a primitive newsreader was chosen as the most flexible option. Making use of the Perl module Net::NNTP (cf. http://www.cpan.org),

this script communicated with the news server of my choice and downloaded a large number of messages as individual files onto my hard disk. In addition, the script interacted with a MySQL database to keep track of what had already been downloaded. In this way, new messages can be downloaded and appended to the corpus at a later stage without any risk of adding duplicate entries.

As a result of this process, a total of 773,772 different messages containing more than 150 million words were downloaded over a period of several days.[11] This figure could easily be increased simply by adding further newsgroups to the list. In Section 4, I will now turn to the procedure required for converting several hundred thousand individual files to a format that can be searched and analyzed with the help of standard corpus tools.

4 Establishing Hierarchical Relationships

Apart from the actual text, each Usenet message also contains a header which lists important information such as the subject of the message, the name and e-mail address of the author and the date and time at which the message was posted to the newsgroup. A sample header is shown in Fig. 2.[12]

As Fig. 2 shows, the header also includes a number of fields whose contents are of a more technical nature. Some of this information is crucial for arranging the individual messages into a meaningful sequence. Most importantly, all messages that are replies to the same topic need to be grouped together—even if the wording of the subject line changes. In addition, it is necessary to determine whether a message is an immediate reply to a new topic or whether it is instead found at the end of a whole chain of replies. For this purpose, messages can be hierarchically organized into so-called threads. A graphical representation of such a thread is shown in Fig. 3. Each circle symbolizes a message which is in some way related to the original post (indicated by a square). Thus, the shaded circle represents a message that is sixteen levels down in the hierarchy of the thread. It is one of three replies to a message found on the fifteenth level and itself has received three further responses. It is not unusual for individual Usenet threads to span across more than 100 levels.

This hierarchical ordering can be achieved with the help of the 'Message-ID' and 'References' elements found in the header. These items are indicated in boldface in Fig. 2. Each message carries a unique identification code such as <MPG.196bfcab94cf0a8989718@news.supernews.net> shown in Fig. 2. If a user writes a response to this message, the new contribution will, of course, also receive a unique 'Message-ID' entry of its own. In addition, the identification code of the message to which the user replies will be appended to the 'References' element in the header. Thus, the message whose header is shown in Fig. 2 is an immediate reply to a message with the identification

```
Path: news5.aus1.giganews.com! nntp4!intern1.nntp.aus1.giganews.com!
From: Dan Swartzendruber <dswartz@druber.com>
Newsgroups: alt.fan.noam-chomsky,soc.history,soc.history.ancient
Subject: Re: What is the most dangerous false belief in the world
today ?
Date: Tue, 1 Jul 2003 21:17:20 -0400
Organization: Posted via Supernews, http://www.supernews.com
Message-ID: <MPG.196bfcab94cf0a8989718@news.supernews.net>
References: <a1333567.0307010632.744e81cd@posting.google.com>
<bdt8ab$ojs$1@newsg4.svr.pol.co.uk>
<MPG.196bf1e0fe42036a989717@news.supernews.net>
<bdtbh7$jhi$1@news6.svr.pol.co.uk>
X-Newsreader: MicroPlanet Gravity v2.50
X-Complaints-To: abuse@supernews.com
Lines: 26
Xref: intern1.nntp.aus1.giganews.com alt.fan.noam-chomsky:132760
soc.history:123006 soc.history.ancient:262349
```

Fig. 2 A sample Usenet header

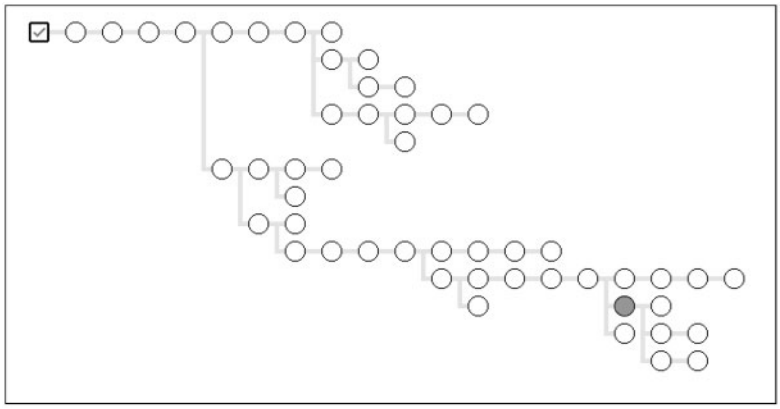

Fig. 3 Graphical representation of a hierarchically organized thread of Usenet messages

code <bdtbh7jhi1@news6.svr.pol.co.uk>. Given this information contained in the header, it is possible to write a simple Perl script which establishes the hierarchical status of each message in the downloaded data. This script first detects all messages without a 'References' element—i.e. messages which start a new topic—and then searches for all messages whose 'References' element contain only the identification code of a root message. These messages are consequently coded as belonging to the first level in the hierarchy of their threads. Furthermore, an additional header element is created which contains the message identification code of the corresponding root message. A similar procedure is then applied to determine all replies to first level posts and to code these contributions as second level messages. The only difference is that the 'References' elements of second—and higher—level messages typically contain more than one identification code from earlier messages. Since additions to the 'References' element are always appended to the right of the existing items, a regular expression match is therefore first required to determine the most recently added identification code. These steps are repeated for successive levels of the hierarchy until no more matches between identification codes in the 'Message-ID' and 'References' elements can be established.

This procedure was applied to all of the 773,772 messages in my downloaded data. However, even though each individual message could now be placed reliably in relation to all the other contributions of the same thread, the nature of the data still required a considerable additional post-processing effort in order to form the basis for investigations into the interactive nature of Usenet discussions. In the following section, I will turn my attention to the text section of the individual messages.

5 Establishing Authorship

As mentioned in Section 2, it is possible for authors of Usenet articles to quote stretches of text from messages to which they reply. Since these old elements can themselves contain previously quoted text, a single Usenet message may in fact consist of (partial) contributions from a potentially large number of different users. This is well exemplified in the article shown in Fig. 4, which contains elements from five different levels of hierarchy in the thread.

As pointed out earlier, in order to be able to carry out a quantitative investigation of the interactive strategies employed on Usenet, the data has to be available in a format which allows researchers to keep track of the authorship of individual stretches of text. Unfortunately, Usenet articles for various reasons often contain a considerable number of formal inconsistencies which make it difficult to reliably establish authorship.[13] In the following,

```
In article <bdtbh7$jhi$1@news6.svr.pol.co.uk>,
agamemnon@hello.to.NO_SPAM says...
>
> "Dan Swartzendruber" <dswartz@druber.com> wrote in message
> news:MPG.196bf1e0fe42036a989717@news.supernews.net...
>> In article <bdt8ab$ojs$1@newsg4.svr.pol.co.uk>,
>> agamemnon@hello.to.NO_SPAM says...
>>>
>>> "Zardoz" <zardoz07@myfastmail.com> wrote in message
>>> news:a1333567.0307010632.744e81cd@posting.google.com...
>>>> What is, in your opinion, the most influential and dangerous
>>>> false belief in today's world?
>>>>
>>>> By false belief, I mean something that had been refuted by the
>>>> experts beyoun a reasonable doubt, but is still held by
>>>> the general public (or a part thereof) as true.
>>>
>>> Neo-Conservatism, Zionism, and Islam.
>>
>> I guess Marxism is not a valid choice, since the "still held by
>> the general public" is no longer true :)
>
> Marx was a Zionist.

Which is irrelevant, as far as I can tell.  Certainly 99.999% of the
people who purported to follow Marxism weren't.
```

Fig. 4 A sample Usenet message with several levels of quoted text

I will concentrate on two of these issues in more detail.

The first area of inconsistency is constituted by the fact that the angle bracket is not a mandatory indicator of quoted text. As a result, some newsreaders give users an option to select a quotation marker of their own choice. Instead of an angle bracket ('>') some messages may therefore for example contain hashes ('#'), vertical dashes ('|') or exclamation marks—or in fact any combination of these—at the beginning of each quoted line. I attempted to solve this problem by writing a simple set of heuristics into my Perl conversion scripts that would interpret recurrent non-word characters at the beginning of lines as quotation markers. However, it soon emerged that such a simple approach would introduce new inconsistencies which would at least partly cancel out the success of my method. For example, in some of the more technically oriented newsgroups, short extracts of programming code often form an integral part of messages. Since some programming languages use hashes to indicate comments, some of these lines of code would, as a result, be interpreted as quoted text. A quick scan of 1,000 randomly chosen messages then revealed that the overwhelming majority of quoted text was indicated with the help of angle brackets. In the end, I therefore decided to disregard other quotation markers and to restrict my algorithms exclusively to angle brackets. Future optimizations of the code may of course lead to a more reliable implementation of the detection of other quotation markers.

The second area of formal inconsistency to be discussed here introduces a potentially much greater level of distortion to the data. It has to do with the fact that the maximum number of characters per line can differ from newsreader to newsreader. Thus, a user may write a message with a newsreader whose line-wrap is set to seventy-four characters. If this message is replied to (and quoted) by another user whose newsreader is set to a shorter maximum line length, parts of each line of the original message will necessarily have to be placed on a second line.[14] Unfortunately, newsreaders are typically not programmed to implement this line-wrap in a consistent way. As a result of this, text that was in reality produced by a single author will normally be split up and displayed as stretches of text that appear to have originated from two different authors. This is best demonstrated with the help of the constructed example shown in Fig. 5.

Original line:

```
>>> This is a sample text which originally appeared on a single line.
```

Ideal format of quoted text after line-wrap:

```
>>>> This is a sample text which originally appeared on a single
>>>> line.
```

Typical output of line-wrap #1:

```
>>>> This is a sample text which originally appeared on a single
> line.
```

Typical output of line-wrap #2:

```
>>>> This is a sample text which originally appeared on a single
line.
```

Fig. 5 Faulty line-wrap of quoted text

```
>>>>>>>>>>>>>>>Are you giving 'props' to Gary?
>>>>>>>>>>>>>>>Yo!!! Give Gary Burnore a chance to defend himself! I'm
sure
>>>>>he'll be
>>>>>>>>>>>>>>>back soon. He's probably out sniffing bicycle seats!!!!
>>>>>>>>>>>>>>>:-)
```

Fig. 6 An authentic example of faulty line-wrap

The sample text displayed in Fig. 5 has three angle brackets at the beginning of the line. In other words, the message originated four levels up in the hierarchy of the thread and was quoted by three successive authors. If this line is quoted for the fourth time and as a result of this exceeds the maximum possible number of characters per line, the ideal solution would be that the newsreader automatically adds only one angle bracket to the first line but a total of four angle brackets to the second line. In this way, both lines of text would still be marked as belonging to the same hierarchical level—and therefore to the same author. In principle, this would be very easy to implement. However, none of the newsreaders I tested opts for this solution. Instead, lines are wrapped as shown in the two typical output samples displayed in Fig. 5. In both cases, the word *line* now looks as if it belonged to a level in the hierarchy of the thread which is different from the original text.

While human readers will have little or no difficulties in linking such separated stretches of text, this situation inevitably results in faulty interpretations when the data is analyzed by means of an automated procedure. As a case in point, consider the short extract shown in Fig. 6. Here, the sentence *I'm sure he'll be back soon* is displayed as if its individual components had been contributed by a total of three different authors. Furthermore, the apparent hierarchical 'distance' between the word *sure* and the actual message to which it belongs spans over fourteen levels of the thread. Since this type of faulty line-wrap occurs fairly frequently, I needed to write a conversion Perl script which would keep track of individual contributions and annotate the actual messages accordingly.

For this purpose, I devised a simple but effective set of algorithms which radically improves the quality of the data. In a first step, I made use of the process described in Section 4 to isolate

messages which have no 'References' element in the header and therefore start a new topic. The text of these 'root messages' was then compared with the contents of corresponding messages which are one level down in the hierarchy and which thus constitute immediate replies. If the reply contained any quoted text, my script first discarded all angle brackets and then attempted to match each line of quoted text with the contents of the relevant root message. A match was only established if every single character of the quoted line (including whitespace characters) exactly corresponded to what was found in the root message.[15] In the majority of cases, such direct matches could be found and the quoted line of text was as a consequence annotated with the identification code of the original message.[16]

However, the authorship of a sizeable proportion of quoted elements could not be ascertained in this first step. One reason for this was that stretches of otherwise identical text were distributed differently across individual lines. I therefore adapted the relevant regular expression to allow for an optional line break to occur after each single word.[17] While the sequence of lexical items in the quoted text thus still had to be identical, differences in the layout of messages no longer had an adverse effect on the procedure.

While this simple change in the regular expression greatly improved the precision of my conversion method, the authorship of a considerable number of quoted text lines could still not be reliably established. In order to increase the overall success rate of my procedure, I therefore introduced a number of additional optimizations. It would go beyond the scope of this paper to present all of these in detail. The following short list must therefore suffice to convey a general impression of my approach.

- When a user shortens quoted text by deleting whole lines or a selection of words in a line, some newsreaders automatically insert elements such as '<snip>' or '[. . .]' to indicate such an omission. This is problematic for my matching algorithms because a line of (partially) quoted text may contain an element which is not present in the original. My regular expression was therefore adapted to disregard a range of typical omission fillers.
- Some newsreaders are not correctly configured to understand line break characters in text documents that were produced on a different operating system. As a result, some newsreaders attach sequences of characters (e.g. '=20') to each individual line. For the same reason, other programs appear to cut off one character at the end of some lines. Again, the regular expression was adapted to account for these types of formal inconsistencies.[18]
- Some users correct spelling mistakes in quoted text. While it could be argued that this in fact improves the overall quality of the data, it of course seriously hampers my efforts of matching quoted elements with their corresponding originals. I experimented with a relatively complex algorithm that would allow a certain percentage of characters to be different in every line of quoted text. However, this was not unproblematic and too often resulted in wrongly matched elements. In the end, I decided to allow a single character to be different per line, but restricted this option to lines containing at least two lexical items.

The list of formal inconsistencies is unfortunately much longer and a corpus compiler could easily spend weeks in the course of this optimization process. However, as the example of the corrected spelling mistakes has shown, there is always the danger that an apparent optimization in the matching algorithm in fact introduces new errors and thus results in a reduction of the overall precision. Given the inherent 'messiness' of Usenet data, it is simply impossible to devise an automated conversion process which results in a version of the data where the authorship of all elements is fully established. Quoted material which could not be successfully assigned was annotated accordingly.

Once all first-level messages—i.e. immediate replies to root messages—were annotated for authorship, my script moved on to the next level of the individual threads and attempted to match quoted elements in these messages to the converted version of the first-level messages. This was done successively with all levels of the hierarchy

```
<header>
     Group:      alt.fan.noam-chomsky
     Message-ID: <MPG.196bfcab94cf0a8989718@news.supernews.net>
     From:       Dan Swartzendruber <dswartz@druber.com>
     Subject:    Re: What is the most dangerous false belief in the
                 world today ?
     Date:       Tue, 1 Jul 2003 21:17:20 -0400
     Root MsgID: <a1333567.0307010632.744e81cd@posting.google.com>
     Level:      4
</header>

<body>
     <4_MPG.196bfcab94cf0a8989718@news.supernews.net> In article
         <bdtbh7$jhi$1@news6.svr.pol.co.uk>,
         agamemnon@hello.to.NO_SPAM says...
     <3_bdtbh7$jhi$1@news6.svr.pol.co.uk> "Dan Swartzendruber"
         <dswartz@druber.com> wrote in message
         news:MPG.196bf1e0fe42036a989717@news.supernews.net...
     <2_MPG.196bf1e0fe42036a989717@news.supernews.net> In article
         <bdt8ab$ojs$1@newsg4.svr.pol.co.uk>,
         agamemnon@hello.to.NO_SPAM says...
     <1_bdt8ab$ojs$1@newsg4.svr.pol.co.uk>"Zardoz"
         <zardoz07@myfastmail.com> wrote in message
         news:a1333567.0307010632.744e81cd@posting.google.com...
     <0_a1333567.0307010632.744e81cd@posting.google.com> What is, in
         your opinion, the most influential and dangerous false
         belief in today's world? By false belief, I mean something
         that had been refuted by the experts beyoun a reasonable
         doubt, but is still held by the general public (or a part
         thereof) as true.
     <1_bdt8ab$ojs$1@newsg4.svr.pol.co.uk> Neo-Conservatism, Zionism,
         and Islam.
     <2_MPG.196bf1e0fe42036a989717@news.supernews.net> I guess
         Marxism is not a valid choice, since the "still held by the
         general public" is no longer true :)
     <3_bdtbh7$jhi$1@news6.svr.pol.co.uk> Marx was a Zionist.
     <4_MPG.196bfcab94cf0a8989718@news.supernews.net> Which is
         irrelevant, as far as I can tell. Certainly 99.999% of the
         people who purported to follow Marxism weren't.
</body>
```

Fig. 7 A converted Usenet message from alt.fan.noam-chomsky

until the complete set of 773,772 messages was available in an annotated format. Figure 7 displays a typical Usenet message from the newsgroup alt.fan.noam-chomsky after this conversion process. The original versions of the header and body of this message have already been shown in Figs 2 and 4, respectively.

The information shown in the header has been considerably reduced. Apart from the fields 'Group', 'From', 'Subject', 'Date', and 'Message-ID', it also displays the identification code of the message which started the thread ('Root MsgID'). In addition, the field 'Level' indicates the hierarchical position of the converted message in the thread. In the body of the message, each line starts with an annotation tag in angle brackets which specifies both the hierarchical level of the message in which the stretch of text was initially produced as well as its unique message identification code.[19]

Of course, the format shown in Fig. 7 may need to undergo further processes of conversion in order to fulfill the specific requirements of a particular task or corpus tool. For example, scholars may want to employ an automatic tagger to annotate the corpus with part-of-speech information. Also, it may be beneficial to convert the data to a fully TEI-conformant XML structure. Since the data now exists in a much more ordered and consistent format, additional conversion steps will involve much less programming effort than the correct assignment of authorship required.

6 Evaluation of the Conversion Procedure

The approach I have outlined in the previous section admittedly does not constitute state-of-the-art programming. For example, there would no doubt be much more sophisticated ways of implementing the method of near-duplicate detection. However, although it uses fairly simple algorithms, my conversion process has produced a corpus which is a great deal more suitable for the linguistic analysis of newsgroup-specific language use than its original format. This claim is supported by Table 2, which offers information about the proportion of successfully assigned quoted material for each newsgroup.

Apart from the total number of articles downloaded for each newsgroup, Table 2 also indicates the number of messages which were found to contain quoted material (i.e. lines which started with an angle bracket). The last two columns on the right refer to the percentage of this quoted material for which authorship could not be determined. The first of these columns shows the proportion of all messages which contain any such material of unknown origin. Interestingly, the values range from a low 2.6% to a proportion that is more than ten times higher: 26.9%. I will offer some further comments on these rather striking differences below. The column furthest to the right, finally, is a variant of the previous one in that for each of the hierarchical levels of a thread only the percentage of newly quoted elements is taken into account. In other words, if an unassigned element is repeatedly quoted on different levels, it is only counted once. In this column, the difference between the highest (17.6%) and the lowest figure (1.4%) is even more pronounced. On the whole, however, given the inherently messy nature of the raw data, the conversion process is certainly successful: for most newsgroups, authorship is correctly assigned for well over 90% of all quoted material.

Since the same conversion procedure was applied to all of the newsgroups listed in Table 2, the considerable variation in the success rates can be interpreted as a reflection of the different strategies that are employed by users when they integrate quoted material into their contributions. The low success rate of my algorithms for the newsgroup rec.gambling.sports, for example, can be explained by the fact that many of the articles that are posted refer to bets that their authors have placed (or will place): many of the root messages mainly consist of listings of sporting events and their projected outcomes. Once the definite results are known, the relevant parts of the original message are quoted in a follow-up message and amended to reflect the actual outcome of the event.[20] Any newly added—and thus unquoted—elements of the message then typically comment on these results. In this particular

Table 2 Success rates for quote assignment in individual newsgroups (proportion of messages with quoted material overall and per hierarchical level)

Name	Messages	With quoted material	% Unassigned	% Unassigned per level
rec.gambling.sports	34,848	17,628	26.9	17.2
alt.fan.noam-chomsky	62,703	52,801	10.9	6.1
alt.alien.research	58,189	48,189	18.3	6
rec.music.classical.recordings	147,866	119,269	8.4	5.1
rec.sport.swimming	15,980	12,799	6.9	4.4
rec.photo.digital	279,846	234,497	6.1	3.6
news.software.readers	31,874	26,480	5.5	3.5
rec.audio.tech	24,645	18,868	5	3.2
alt.music.oasis	21,883	14,989	4.7	3.1
alt.coffee	68,056	50,243	4.4	3
news.software.nntp	5,033	3,709	4.6	2.9
alt.games.warcraft	23,849	20,135	2.6	1.4
Total	773,772	619,607		

newsgroup, the boundaries between old and new material are therefore blurred. Since my matching algorithms are not designed to account for this type of use, a large proportion of the quoted material of course cannot be properly assigned.

Another interesting pattern is found for the newsgroup alt.alien.research, where the authorship of over 18% of all messages could not be properly assigned. However, if only newly quoted material is considered, this figure drops to a much lower 6% of all messages. A detailed analysis of these figures is beyond the scope of this work, but this difference suggests that users of this newsgroup tend to quote individual stretches of text over many levels of the hierarchy of the thread.

The data contained in Table 2 provides an interesting perspective on the nature of written communication in the various newsgroups. Thus, it would clearly be wrong to consider the language of Usenet as a homogeneous phenomenon. Instead, it is a much more likely scenario that the individual newsgroups mirror the diversity of spoken interaction and that different genres of written communication also exist on Usenet. As a result of the relatively simple conversion process described in this article, these differences are now much more easily accessible for quantitative linguistic research.

7 Conclusion

In this article, I have presented an outline of the procedure required to convert a large collection of messages that were downloaded from twelve Usenet discussion groups into a format which allows researchers to investigate the communicative strategies that are encountered in this specific type of written communication. In particular, this required that the authorship of quoted material, which forms an integral part of Usenet messages, be correctly assigned. For this purpose, it was first necessary to determine the hierarchical ordering of the messages into individual threads. In a second step, I then proceeded to match the contents of messages found on adjacent levels of the hierarchy. This was initially done on a line-by-line basis, but a number of refinements had to be introduced to account for the various types of distortions that are introduced by the less-than-optimal way in which newsreaders typically handle quoted material. While it was not possible to eliminate these inconsistencies completely, my matching algorithms had a success rate of well over 90% for most of the newsgroups that were considered for this study.

As I mentioned in the introduction, only fairly basic programming skills are required for carrying out this type of conversion process. I hope that readers may therefore now feel encouraged to create their own specialized Internet-derived corpora. In the absence of a reliable linguistic Web search engine, such an approach greatly improves our possibilities of investigating language use in a medium which has already considerably changed the way we communicate.

However, an important question remains. To what extent can Internet-derived corpora complement or perhaps even replace existing language corpora? In this context, it needs to be stressed that even after the clean-up process described in the present work, the collection of Usenet data remains very messy in several respects:

- There is very little reliable information available about the authors of Usenet messages. Since spam is a serious issue, many posters do not reveal their correct e-mail addresses and affiliations but instead either artificially distort them or just invent mock identities (e.g. god@earth.com). This type of information can easily be configured in the preferences settings of the newsreader and it is at least theoretically possible to change these settings for every message. As a result, messages that are written by the same person may look as if they were posted by completely different authors.[21] On the whole, researchers need to accept the fact that next to no information is available that could form a reliable basis for any sort of sociolinguistic investigation of the data.
- Secondly, even if creators of Usenet messages could be identified, the actual authorship of individual text elements may still remain unknown. For example, it is a standard procedure to copy whole stretches of text that

were found on a Web page and integrate them into Usenet messages. Apart from the resulting difficulties in assigning authorship, this of course also means that the degree of orality in Usenet messages can vary a lot. In fact, in many newsgroups the language use encountered typically ranges from highly formal and edited written text to very speech-like, short 'utterances' resembling oral communication.

- A third source of potential distortion of the data originates from users who flood Usenet discussions with advertisements and other types of spam. Their messages typically have little or no connection with the topic of the newsgroup and their contributions are of course not intended as genuine subject matter for discussion.

This incomplete list of shortcomings suggests that researchers must proceed with great caution when Internet-derived data is employed as a proxy for general language use. Indeed, for the purpose of most research questions, such type of data clearly cannot replace the carefully compiled electronic language corpora available today. However, this should certainly not deter anybody from tapping into the wealth of data available on the Internet. For example, given the comparative ease with which very large collections of authentic language data can be compiled, such Internet-derived corpora can form an important additional basis for investigations of phenomena whose frequency is so low that even corpora of the size of the 100-million word British National Corpus cannot supply a sufficient number of instances. Furthermore, Usenet offers researchers direct access to very recent language data and may therefore provide a suitable basis for an analysis of ongoing changes in the system of communication. In sum, Internet-derived corpora clearly are highly valuable collections of authentic language use which complement currently available electronic language corpora.

References

Baroni, M. and Bernardini, S. (eds) (2006). *WaCky! Working Papers on the Web as Corpus*. Bologna: Gedit.

Baym, N. K. (1996). Agreements and disagreements in a computer-mediated discussion. *Research on Language and Social Interaction*, **29**(4): 315–45.

Buchanan, L. (2000). Performing gender online: extending the search for male and female electronic message variants. *Southern Journal of Linguistics*, **24**(2): 147–63.

Burgess, C. (2001). Representing and Resolving Semantic Ambiguity: A Contribution from High-Dimensional Memory Modeling. In Gorfein, D.S. (ed.) *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*. Washington, DC: American Psychological Assoc, pp. 233–60.

Burke, S. M. (2002). *Perl & LWP*. Sebastopol, CA: O'Reilly.

Cheung, C.-S. and Fung, P. (2004). Unsupervised learning of a Chinese spontaneous and colloquial speech lexicon with content and filler phrase classification. *International Journal of Speech Technology*, **7** (2–3): 173–188.

Fletcher, W. (2002). Making the Web a More Useful Source for Corpus Linguistics. In Connor, U. and Upton, T. (eds), *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*. Amsterdam: Rodopi, pp. 191–05.

Galegher, J., Sproull, L. and Kiesler, S. (1998). Legitimacy, authority, and community in electronic support groups. *Written Communication*, **15**(4): 493–530.

Gheno, V. (2003). Prime osservazioni sulla grammatica dei gruppi di discussione telematici di lingua Italiana. *Studi di Grammatica Italiana*, **22**: 267–308. Available at http://www.accademiadellacrusca.it/riviste/riviste.php?vedi_elenco=1&ctg_id=75&pag=3 (accessed 19 February 2007).

Kilgarriff, A. (2003). Linguistic Search Engine. In Archer D., Rayson, P., Wilson A. and McEnery, A. (eds), *Proceedings of the Corpus Linguistic 2003 Conference*. Lancaster: UCREL. 53–58. Also available at ftp://ftp.itri.bton.ac.uk/reports/ITRI-03-19.pdf (accessed 5 December 2005).

Lewin, B. A. and Donner, Y. (2002). Communication in Internet message boards. *English Today*, **18**(3): 29–37.

Mair, Chr. (2007). Change and Variation in Present-Day English: Integrating the Analysis of Closed Corpora and Web-Based Monitoring. In Hundt, M., Nesselhauf, N. and Biewer, C. (eds), *Corpus Linguistics and the Web*. Amsterdam: Rodopi, pp. 233–48.

Mair, Chr. (2003). Tracking Ongoing Grammatical Change and Recent Diversification in Present-Day Standard English: The Complementary Role of Small and Large Corpora. Paper presented at the annual ICAME conference, Guernsey. http://fips.igl.uni-freiburg.de/lsf/docs/Mair.pdf (accessed 4 December 2005).

Marcoccia, M. (2004). On-line polylogues: conversation structure and participation framework in Internet newsgroups. *Journal of Pragmatics*, **36**: 115–45.

Olavarría de Ersson, E. and Shaw, P. (2003). Verb complementation patterns in Indian Standard English. *English World-Wide*, **24**(2): 137–61.

Renouf, A, Kehoe A. and Banerjee, J. (2005) The WebCorp Search Engine: A Holistic Approach to Web Text Search. In *Proceedings of Corpus Linguistics 2005* Birmingham, 14–17 July 2005, University of Birmingham.

Rohdenburg, G. (2007). Determinants of Grammatical Variation in English and the Formation/Confirmation of Linguistic Hypotheses by Means of Internet Data. In Hundt, M. Nesselhauf, N. and Biewer, C. (eds), *Corpus Linguistics and the WEB*. Amsterdam: Rodopi, pp. 191–210.

Thelwall, M. (2005). Creating and using Web corpora. *International Journal of Corpus Linguistics*, **10**(4): 517–41.

Wall, L., Christiansen, T. and Orwant, J. (2000). *Programming Perl*, 3rd edn. Sebastopol, CA: O'Reilly.

Werry, Chr. C. (1996). Linguistic and Interactional Features of Internet Relay Chat. In Herring, S. C. (ed.) *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam: John Benjamins, pp. 47–63.

Notes

1 For a more detailed discussions of these options, see Thelwall (2005).

2 Some of these drawbacks are alleviated by linguistically oriented front-ends to commercial search engines such as WebCorp (Renouf *et al.*, 2005) or KWiCFinder (Fletcher, 2002). However, while their output format is certainly more appropriate for linguistic research, their reliance on commercial search engines results in the same kind of general search limitations that are imposed by direct Google or Altavista searches.

3 Also, commercial online databases such as those provided by LexisNexis (http://www.lexisnexis.com) and Chadwyck-Healey (http://collections.chadwyck.co.uk/) contain a wealth of linguistically interesting data in a searchable format.

4 A good example of the use of online newspaper archives is Olavaría de Ersson and Shaw's (2003) investigation of verb complementation patterns in British and Indian English. In addition to the presentation of linguistic results, the authors also provide a discussion of the methodological issues involved when newspaper online search facilities are used.

5 A comprehensive introduction to programming with Perl is given in Wall *et al.* (2000). Readers may also wish to consult Burke (2002) for an in-depth overview of regular expressions.

6 Readers who are interested in examining the Perl scripts written for the procedure described in this article are invited to contact me directly.

7 Nowadays, however, the existence of many Web-based newsreading services blurs the boundaries between Usenet and the World Wide Web.

8 In contrast to a typical Web page, a Usenet message is not stored on a single server; instead, it is distributed to all news servers which are configured to host the particular newsgroup to which the message was posted. When users in different parts of the world read a particular message, they will thus access a local copy which was delivered as part of the 'newsfeed' to the news server of their choice. Since different servers have different settings for updating newsfeeds, a situation may arise where users can in fact access a reply to a message before they can read the original post.

9 Since news server have limited capacities, messages are set to expire after a certain time and are then deleted from the local servers to make room for new posts. This time—which is also referred to as 'retention rate'—differs from server to server. As a result, newsgroup discussions which are conducted over a period of several days or weeks may be incomplete on some servers while they are still fully accessible on others. Discussions that have been deleted from all news servers can usually still be accessed via the Google Usenet archive, which contains messages posted as early as May 1981 (cf. http://groups.google.com).

10 In the latter case, the retention rate of the news server from which messages are downloaded will be an important factor to consider. While some newsgroups generate dozens or even hundreds of messages every day, a short expiration period may mean that only a small number of messages will be available for many of the less popular groups. As a general rule,

commercial news servers appear to offer significantly longer retention rates than those provided by educational institutions. A list of commercial services can be found at <http://dmoz.org/Computers/Usenet/Feed_Services/>.

11 The actual download process is relatively slow since each message has to be accessed individually. Speed can vary greatly depending on the news server used and the number of concurrent connections that can be established. A typical download rate can lie anywhere between one and ten messages per second.

12 Most newsreaders hide at least some of this information from users as its content is likely to distract from the actual message.

13 With the term *formal inconsistency*, I am referring to both differences in the layout of Usenet messages as well as the use of different formal features for the same purpose.

14 Typically, newsreaders account for the fact that each time a line is a quoted, it will receive one or two extra characters (an angle bracket and often a whitespace). Thus, as a precaution, original messages are usually line-wrapped at a lower character count than necessary. However, if a line is quoted several times in succession, the total character count of the line will soon reach the internal limit, which then results in potentially erratic line-wrap behavior.

15 This is a slightly simplified account of the actual procedure. For example, the script remembers the position of a previously matched quoted line and restricts new searches to those portions of the root message which follow this position. Difficulties would otherwise arise with short lines that consist of frequently used lexical material and which match several times in the same message.

16 Those lines of the message which were not marked as quoted material and thus constituted new elements of text were also annotated for authorship and received a tag with the identification code of the current message.

17 For this purpose, each whitespace character in the line of quoted text was replaced with an optional line break character that is surrounded by optional whitespace characters ('\s*\n*\s*'). In this way, the regular expression still matches any single whitespace between two words but also matches newly introduced line breaks regardless of whether they precede or follow a whitespace.

18 Readers may be surprised at the range of basic inconsistencies introduced by some newsreaders. This may have to do with the fact that it is a relatively simple task to write programming code for a basic newsreader, and many 'home-made' variants appear to be in circulation which clearly do not meet the standards of full-fledged commercial software tools.

19 In Fig. 7, most textual elements span across several lines. This layout is for display purposes only and therefore does not reflect the actual formatting.

20 This 'reply' is often contributed by the same author as the original message, which adds to the rather unusual communicative patterns found in this newsgroup.

21 A second possible source of information about authors is found in the signatures that are often appended to messages. But here again, this is very unreliable data.