

Chapter 4. Findings

4.1. Introduction

This study sought to 1) determine how question type affects the actions performed by triagers, and 2) draw up a set of rules for the performance of triage based on the question type being triaged. Question type was defined as the three classes into which a question was classified, according to three taxonomies of questions.

At the beginning of the study, four taxonomies of questions were identified in the literature from desk and digital reference, question answering, and linguistics. These four taxonomies correspond to the top four levels of linguistic analysis. This study classified questions according to three of these four taxonomies: the subject taxonomy was not utilized because it is a matter of preference which of the large number of available subject classification schemes is used, and different digital reference services use different schemes. (See section 3.3 for a detailed discussion of why questions were not classified according to subject.) These taxonomies were evaluated according to a set of thirteen criteria, and modified accordingly.

Attributes of questions that affect the triage process were determined by observing triagers performing the task of triage, utilizing the think-aloud methodology. Eight attributes of questions (and a total of thirty-eight criteria of different types) were discovered that affect the triage process.

Question type was determined by classifying questions according to all three taxonomies. Nine coders classified questions, and the intercoder reliability statistic, Cohen's κ , was computed between the coders' classifications (see section 3.3.4 for a detailed discussion of κ). The values of κ ranged from 0.53 to 1, indicating good to perfect reliability. The correlation between question type and the action taken on a question in the triage process was determined by calculating Cramér's V . The values for Cramér's V ranged from 0.06 to 0.27, indicating a weak correlation across all classes in the three taxonomies. A few specific question types in the "taxonomy space" defined by these three taxonomies –

intersections of classes within the three taxonomies – correlated very strongly with the action taken on a question in the triage process.

This chapter discusses in detail the findings of this study, in order of the steps in the study methodology, as presented in Figure 4-1.

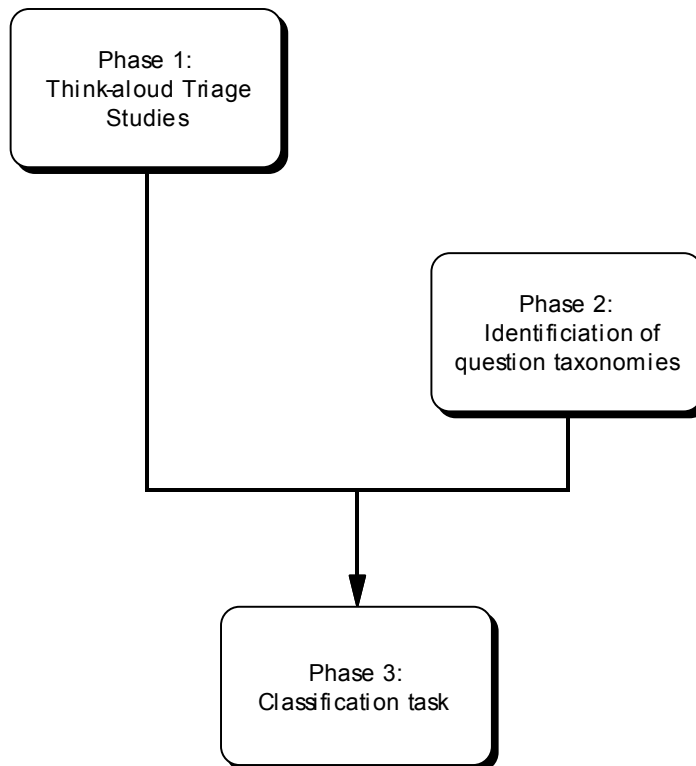


Figure 4-1: Research Methodology

4.2. Phase 1: Think-aloud Studies

The think-aloud phase of this study addressed Research Question 1, and its two sub-questions:

RQ1. What attributes of questions affect the triage process?

RQ1a. What attributes of questions are taken into account by digital reference triagers when performing triage on received questions?

RQ1b. How do these attributes affect triagers' decisions in triaging questions?

This section discusses the findings from the think-aloud phase of this study, and answers Research Questions 1 and its two sub-questions. A total of thirty-eight criteria that influence triage decisions were discovered, eight of which were intrinsic to the question itself; thirty criteria were extrinsic to the question, situating it in a context for the user and the service.

4.2.1. Participating Digital Reference Services and Triagers

A pool of 175 potential services was solicited for participation in this study. Of these, twenty-eight services participated in the think-aloud studies, performed between August and October 2002. In other words, twenty-eight triagers from twenty-eight different digital reference services were studied in this phase of the study. The break-down of these twenty-eight services by type is as follows:

- Academic libraries: 57.1%
- AskA services: 21.4%
- Public libraries: 10.7%
- Special libraries: 10.7%

Additionally, the services solicited are located in several countries. The break-down of the locations of these services is as follows:

- United States: 84.8%
- Canada: 6.1%
- Australia: 3.0%
- Netherlands: 3.0%
- United Kingdom: 3.0%

Of these participating services, 85% maintained web-based question submission forms, and 15% did not. Of the services that did not maintain web-based question submission

forms, all published one or more email addresses on their websites to which questions could be submitted. At least one of these services published a list of email addresses for the various departments in the library to which reference questions could be sent (e.g., Archives, Government publications, Interlibrary loan, Special collections, etc.), and at least one of these services published a list of email addresses for the various subject specialists employed by the library to whom reference questions could be sent (e.g., American literature, Asian studies, Chemistry, Mathematics, etc.).

4.2.1.1. Reasons for Services' Non-Participation in this Study

Out of the pool of 175 potential respondents solicited for participation in this study, twenty-eight services responded to the solicitation, and participated in the think-aloud studies. This section discusses the characteristics of those 142 services that were solicited but did not participate in this study; these services are referred to as non-participating services. The break-down of these twenty-eight services by type is as follows:

- Academic libraries: 28.6%
- AskA services: 42.9%
- Public libraries: 25.9%
- Special libraries: 2.7%

The break-down of the locations of these services is as follows:

- United States: 83.7%
- Canada: 12.2%
- Australia: 1.4%
- Singapore: 1.4%
- Netherlands: 0.7%
- United Kingdom: 0.7%

The break-down of the locations of participating and non-participating services is approximately equal. The break-down of the type of participating and non-participating services is not equal: the pool of services that participated in this study contains a greater

percentage of academic and public libraries, and a smaller percentage of AskA services and special libraries than the pool of services that did not participate. This finding points to a possible sampling bias, in that AskA services and special libraries may be over-represented in the pool of services that participated in this study. It is worth pointing out, however, that the unit of analysis throughout this study is the question, not the service, so it is not clear that this sampling bias would have had any impact on the distribution of question types represented in the pool of questions collected in this phase of the study.

There were a total of sixteen reasons that the 142 non-participating services did not participate in the study, presented in Table 4-1:

Table 4-1: Reasons for Services' Not Participating in the Study

Non-solicitation Reason		Percentage of the Pool of Potential Respondents
Technical difficulties	Login required	5.6
	Could not connect to the server	1.4
	Question submission form would not submit	0.7
Problems contacting the service	Service never replied to the solicitation	38.7
	No contact information available	2.8
	Duplicate service	1.4
	Service discontinued	1.4
Inappropriate forms of reference offered	Previously-answered question archive only	2.1
	Chat only service	0.7
Inappropriate forms of triage, or triage	A single expert answered all questions	19.0
	Triage by subject only	7.7

not performed	Triage by subject or department only	6.3
	Experts self-selected questions	4.2
	Triage by department or branch only	4.2
	Patron self-triages	2.1
	Questions triaged equally among experts	1.4

There were a number of reasons why services did not participate in this study that had nothing to do with triage specifically. These were the first three categories of reasons listed above: technical difficulties, problems contacting the service, and inappropriate forms of reference offered.

If technical difficulties were encountered, the digital reference service was never contacted and therefore never solicited. If the researcher could not connect to the service's server, or if the web question submission form would not submit (and no email address was listed as an alternative to the webform), an attempt was made again the following week, and then again the following week, so that in total three attempts were made to contact the service, over the span of three weeks. This was done because it is common on the Web for sites to go offline temporarily for any number of reasons, so it seemed unfair to eliminate services from this study because their downtime just happened to correspond with the timeframe of this study. If technical difficulties were encountered on the third attempt, this was taken to indicate a more serious technical problem, and the service was abandoned for this study. On the other hand, a few services required that the patron log in using a student ID or a library card number in order to submit a question to the service. In this way, these services could insure that they were only serving their primary patron community: the universities' faculty, staff, and students, or the city's residents. When alternative contact information was available, these services were solicited by email or phone, rather than *via* the service's website. When alternative contact information was *not* available, these services were abandoned for this study, as

this was an insoluble problem: if the researcher was not affiliated with a service by dint of not being affiliated with a particular institution or not being a resident of a particular geographic area, there was nothing that could be done about that.

A small percentage of services sampled from the LIBWEB database and AskA Locator were duplicates – for example, NASA maintains several AskA services (Ask Dr. SOHO, Mars Team Online, Ask A High Energy Astronomer), all of which employ different pools of subject experts to answer questions, but all of which are managed by the same group at NASA – therefore, the triage process is performed identically. A similarly small percentage of services had been discontinued, for reasons not specified on their webpages. Additionally, on some services' websites there was no information available concerning how to submit a question either *via* a webform or email. It is possible that these services could have been discontinued, or it may simply be that the website did not contain the appropriate information. On the other hand, a remarkably large percentage of services solicited simply never replied.

One University library did not offer email reference at all, and only offered reference online *via* chat. This is particularly remarkable because email reference service has been offered in academic settings since the mid-1980s, but reference service *via* chat application is a comparatively new technology. It is unusual for an academic library (not known to be the most technologically cutting-edge institutions) to abandon an older and better accepted technology – or to not implement it in the first place – in favor of a newer and still experimental one.

A small percentage of non-responding services avoided the issue of triage entirely by requiring or at least requesting the patron to select the library, individual subject specialist librarian, or department to which they wish their question to be sent. This is feasible only for services that are small enough in scope that the patron can easily browse the list of choices and make a decision. Indeed, one service that was located in one library in a University even stated that they have no contact with the reference services in the other libraries on campus and consequently have no idea of those other libraries' policies

and practices regarding digital reference service. Thus, requiring patrons to “self-triage” their questions may ease or eliminate any responsibility for triage by the digital reference service itself, but this may come at the cost of losing collaboration among services. On the other hand, the question can be triaged to a different recipient if the recipient chosen by the patron deems the question inappropriate or out of scope.

Another reason why services did not participate in this study was that triage was not performed at all: either experts self-selected questions or a single expert answered all questions received by a service. The former – experts self-selecting questions – is the method employed by the Internet Public Library (IPL)’s Reference Center using their software application, QRC (Lagace and McClennen, 1998). This is also the method being increasingly employed by the VRD, as it moves its volunteer experts from email to the QABuilder software application. As Pomerantz and others (forthcoming) discovered, there exists “a spectrum of technology use, ranging from highly automated to entirely human-intermediated services” (p. 14), and such web-based and database-driven applications have, to date, only been adopted by more highly automated services. Indeed, the services identified in this study that allow experts to self-select questions are all high-tech services, as well as, incidentally, all being AskA services.

The latter reason why triage was not performed by a service – one librarian or expert answered all questions received on any given day – is feasible only in services that receive a low enough volume of questions that a single individual can answer all questions, as part of their regular set of daily tasks. In fact, all of the services identified in this study that answer questions in this way are extremely low-volume: one service even stated that “we average one e-mail reference question every ten days to two weeks.” Nearly half (44.4%) of the services that answer questions in this way are affiliated with public libraries, and 29.6% of the services that answer questions in this way are AskA services of modest proportions – so modest, in fact, that they are run entirely by one extremely dedicated individual.

For those services affiliated with public and academic libraries (that is, those not run entirely by one person), the fact that one librarian or expert answers all received questions on any given day is actually not entirely incompatible with the performance of triage. Several of these services stated that as a matter of policy the librarian on “email duty” on any given day has primary responsibility for answering reference questions received by email, but if he or she could not answer a question for whatever reason, then he or she could forward it. One service even provided their “Correspondence Policy” on this issue:

Responsibility for replies

Inquiries received by [service’s name] that are general in nature or deal with subjects that fall within the collecting responsibilities of Research Services and Collections’ librarians should in almost all instances be answered by staff on duty at the Reference Desk.

- A. If an inquiry is beyond the scope of Reference Assistants’ responsibilities for service, the librarian on duty or next to come on duty assumes responsibility.
- B. Inquiries are forwarded to an RSC subject specialist only after preliminary consultation with them.

Some triage is therefore inevitably performed even in services in which one librarian has primary responsibility for answering all questions. As stated above, however, all of these services are extremely low-volume. It could therefore have taken many days or even weeks before one question that required triaging was received by a service of this type. Thus, even though triage was occasionally performed in these services, they were eliminated from the pool of potential respondents as requiring an exceptional commitment of time for a very minor return of data. Eliminating these services leads to a sampling bias in that very low-volume services are not represented in the pool of services that participated in this study. Again, however, as the unit of analysis throughout this

study is the question and not the service, it is not clear that this sampling bias would have had any impact on the distribution of question types represented in the pool of questions collected in this phase of the study.

The remaining reason why services did not participate in this study was due to the fact that triage was performed, but according to a strict set of criteria. This involved triage by one of two criteria only: to either the appropriate subject specialist (either an official subject specialist librarian employed by the library or an individual who is informally known to have some specific subject expertise) or to the appropriate library branch or department within the library (circulation, inter-library loan, technical services, special collections, etc.). Again, even though triage was performed in these services, they nevertheless declined to participate in this study, possibly believing that with such simple criteria for triage, they had little to contribute. Even so, these two criteria correspond to two of the most important factors that affect triage (which will be discussed further below): the subject of the question and the subject expertise of the service or answerer to whom the question is triaged.

4.2.1.2. Representativeness of the Sample

Janes (2002) states that “no definitive list of reference librarians exists” (p. 550). Similarly, no definitive list of reference *services* exists, and more importantly for the purposes of this study, no definitive list of digital reference services exists. Thus it is impossible to know what the “demographic” makeup of the population of digital reference services is, and therefore similarly impossible to know what percentage or segment of the total population of digital reference services is made up by the twenty-eight responding services. In spite of this, these twenty-eight services are as representative of the universe of digital reference services in English-speaking nations as it is possible to get. Cook and Campbell state that “the most representative samples will be those that are randomly chosen from the population” (p. 75), so digital reference services were randomly sampled from the sources discussed below. While it is impossible, in the absence of data about the makeup of the population of digital reference

services, to test the actual representativeness of this study's sample of services, this sample is "representative enough."

There are two methods for classifying digital reference services: by the type of library the service is affiliated with, and by the service's use of automation. The former, and the more common method for classifying digital reference services is by the type of library the service is affiliated with: academic, public, or special library. Alternatively, a digital reference service may be unaffiliated with any library, physical or digital; this latter type of digital reference service is referred to an AskA service. The sample of twenty-eight responding services for this phase of the study includes services affiliated with libraries of all types, academic, public, and special, and includes AskA services as well. Additionally, this sample includes services that perform all tasks manually, to highly automated services. Thus, while the makeup of the population of digital reference services is not known, this sample covers as wide a range of different types of services as it is possible to identify in this population.

The LIBWEB database is the most complete database of library websites that exists online as of this writing, listing over 6,500 library websites in over 100 countries, and of all types. By randomly sampling libraries from the LIBWEB database, a "representative enough" sampling of all types of libraries worldwide was ensured for this phase of the study.

The one limitation to the generalizability of the sample for this phase of the study is the fact that only services in English-speaking nations were sampled. This was done for purely practical reasons: the researcher is not fluent in any language but English, and for this phase of this study, it was necessary that the researcher speak to the respondents, if not in person then on the telephone. This fact therefore limits the generalizability of this phase of the study; this study cannot claim to be generalizable to libraries in non-English-speaking nations.

The AskA Locator, like the LIBWEB database is for library websites, is the most complete database of AskA services that exists online as of this writing. By randomly sampling AskA services from the Locator, a “representative enough” sampling of AskA services was ensured for this phase of the study.

As already mentioned, no definitive list of digital reference services exists. It is therefore impossible to know what percentage or segment of the total population of digital reference services is made up by the set of services listed in the LIBWEB database and the AskA Locator. The is the most complete database of library websites that exists online as of this writing, and the AskA Locator is perhaps the only database of AskA services that exists online. The set of services listed in these two sources may not be the entire universe of digital reference services in the world. This set of services is, however, as complete an estimation of that universe as it is possible to get.

The method for classifying digital reference services by the type of library with which the service is affiliated has been used in much of the literature analyzing digital reference services (Janes, Carter, and Memmott, 1999; Garnsey and Powell, 2000; Janes, Hill, and Rolfe, 2001). As mentioned above, the break-down of these twenty-eight services by type is as follows:

- Academic libraries: 57.1%
- AskA services: 21.4%
- Public libraries: 10.7%
- Special libraries: 10.7%

Pomerantz and others (forthcoming), however, make the argument for “a more complex grouping scheme based upon functional, rather than organizational, characteristics” (p. 14). This method for classifying digital reference services is by the service’s use of automation. This classification method was proposed by Pomerantz and others, in a study of the paths digital reference services take through a general process model of asynchronous digital reference. This classification divides digital reference services in three groups: “High Tech/Low Touch,” “Low Tech/High Touch,” and “High Tech/High

Touch.” These names refer to the amount of both automation and human intermediation employed by digital reference services in providing asynchronous digital reference. These three groups are not entirely distinct, but rather are clusters around three points on a spectrum of technology use, ranging from highly automated to entirely human-intermediated services.

Pomerantz and others (forthcoming) list nine key characteristics that differentiate between digital reference services to place them in one of these three groups. In the present study, a short structured interview was conducted with participants after the think-aloud study was conducted, to collect some supplementary data about the service itself and the triage process as performed by the service, if this data did not come up during the think-aloud study or the solicitation exchange. Collecting data on these nine characteristics of digital reference services allowed the twenty-eight responding services to be classified into the three groups named in the previous paragraph. These results are discussed in section 4.2.3.

4.2.2. Think-aloud Task

As mentioned above, twenty-eight think-aloud studies were conducted with triagers from twenty-eight different digital reference services. During these twenty-eight think-aloud studies, 185 questions were triaged and collected. This averages to 6.6 questions per service – though the maximum of 30 questions were triaged during the think-aloud studies with three services, which means that actually there was an average of 3.8 questions per service for the other twenty-five services. Approximately equal numbers of questions were collected from low- and high-volume services.

4.2.2.1. Achieving Saturation

Think-aloud studies were conducted until saturation was achieved. Saturation was achieved early in the think-aloud phase of the study. This is probably mostly to do with the fact that the first two think-aloud studies were performed with respondents from high-

volume services (AskERIC and the Virtual Reference Desk). This was not a deliberate choice, it was an accident of timing due to the fact that because the researcher is affiliated with the Information Institute of Syracuse, which is the umbrella organization that oversees AskERIC and the VRD Project, it took less time to set up the think-aloud studies for these services than for other services, which had to be set up by email and telephone.

Because the first two think-aloud studies were performed with respondents from high-volume services, many reasons for triage decisions were collected right away. There were 60 questions triaged between the first two services (a think-aloud study continued for as many questions as were received by services on that day, up to a maximum of 30 questions), which accounts for nearly a third (32.4%) of the total pool of 185 questions collected. And indeed, saturation occurred when the think-aloud study was approximately 40% completed: with the 11th service (out of twenty-eight), at the 73rd question. After this point no additional factors that affect triage decisions were collected. This “over-saturation” allows for a great deal of confidence that the range of factors that affect triage that were discovered was exhaustive.

4.2.2.2. Volume of Questions Received

The twenty-eight services studied during this phase of the study covered the spectrum of volume of questions received, from very high-volume to very low-volume services. High- and low-volume is a relative measure, as demonstrated by the fact that one service received in excess of 80 questions on the day that the think-aloud was performed (though the think-aloud study only continued for 30 questions), and the respondent claimed that it was a light day, while other participant services routinely receive five or fewer questions per day, so that receiving even ten in one day would be an exceptionally high volume. Indeed, the set of services studied during this phase of the study could be considered to be two separate populations: the populations of high-volume and low-volume services. Data analyses will be presented throughout this chapter, however, to illustrate that the

populations of high-volume and low-volume services do not significantly differ in any characteristic that is relevant to this study, except for the volume of questions received.

One unexpected finding of this study is that the volume of questions received by the participating digital reference services roughly follows a Zipfian distribution. The researcher had anticipated that the volume of questions received by digital reference services would roughly follow a normal distribution: a few services would receive either a very low or a very high volume of questions, and many services would receive a moderate volume. Instead, there were a few services that received a high volume of questions, and many services that received a low volume of questions. Table 4-2 lists and Figure 4-2 shows this distribution. There is, however, an artificial ceiling of 30 on this distribution, since the think-aloud studies only continued for 30 questions per service, and the actual number of questions received by a service, if it was over 30, was not recorded.

Table 4-2: Number of Think-aloud Questions Received on the Day the Think-aloud Study was Performed

Number of questions received	Number of services that received that number of questions	Percentage of services that received that number of questions
30	3	10.7
24	1	3.6
16	1	3.6
14	1	3.6
4	2	7.1
3	4	14.3
2	5	17.9
1	11	39.3

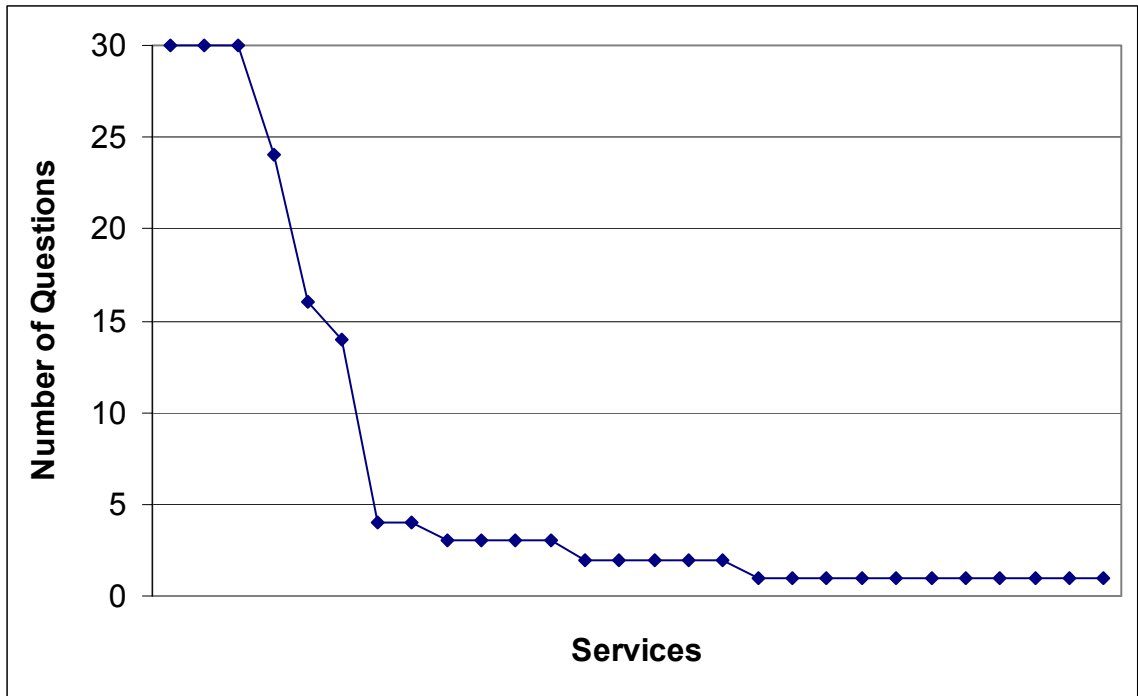


Figure 4-2: Distribution of Number of Questions Received by Participating Services on the Day the Think-aloud Study was Performed

The fact that the volume of questions received by the participating digital reference services was distributed in this way eliminates one potential source of bias in this study. If the number of high- and low-volume services among the respondents were roughly even, or worse, if there were more high- than low-volume services, the majority of questions sampled would be from high-volume services, thus possibly skewing the distribution of question types in the pool. The fact that there were many more low- than high-volume services among the respondents, however, eliminates the possibility of this bias: there were approximately equal numbers of questions collected from the think-aloud studies from low-volume as from high-volume services. Thus, the distribution of questions from high- and low-volume services in the pool should be more approximately even.

That said, there is at present no evidence to support a hypothesis that the types of questions received by high- and low-volume services are different in any way. No studies exist that analyze the types of questions, relative to the volume of questions received by services. Carter and Janes (2001) analyze questions according to a variety of different criteria, but these questions were all received by the Internet Public Library, which is a high-volume service (Carter and Janes report that the IPL received 3,022 questions during the three-month span of their study, an average of approximately 34 questions per day (p. 254)).

4.2.2.3. Factors in Triage Decision-making

All think-aloud studies were tape recorded. The researcher transcribed these recordings to create a written protocol for every respondent, and these protocols were imported into ATLAS/ti, a software application “for the qualitative analysis of large bodies of textual, graphical, audio and video data” (www.atlasti.de/intro.shtml). In ATLAS.ti these protocols were coded so that comments by respondents that indicated reasons for triage decisions were marked in the protocol. These codes were arrived at both deductively and inductively, utilizing the constant comparative method: the fifteen factors that Pomerantz, Nicholson, and Lankes (2003) discovered that influence the triage process served as the initial set of categories into which the attributes of questions, elicited in this think-aloud studies, were coded. As data was collected from the think-aloud studies, more attributes that influence the triage process emerged, and more categories of attributes were developed until saturation was achieved.

A total of eight attributes of questions – attributes intrinsic to the question – were discovered that affect triagers’ decisions on actions to take on questions in the triage process. Attributes of other factors that affect triagers’ decisions were also discovered – attributes extrinsic to the question, that situate it in a context for the user and the service. A total of thirty-eight attributes were discovered in all. These thirty-eight attributes fell into eight categories:

Table 4-3: Categories and Attributes That Affect Triagers' Decisions

Category of Attributes	Attribute
Attributes of the question	Subject of the question
	Difficulty of the question
	Generality or specificity of the question
	Question type, according to one or more of the classes from the taxonomies of wh-words or functions of expected answers
	Interestingness of the question, in the opinion of the triager
	Language that the question is written in
	The fact that one patron submitted multiple questions
The fact that the question has a prior history with the service: it is a follow-up question or has been forwarded to another service and the other service forwarded it back	
Attributes of the answer	Form in which the patron specifies that he or she would like the answer presented: e.g., short or long answer, or according to one or more of the classes from the taxonomy of forms of expected answers
Attributes of the patron	Organizational affiliation
	Role, job description, or other capacity in which the question is being asked
	Age

	School level
	Country of residence
	State of residence, or other subdivision within the country of residence: e.g., Province in Canada
Attributes of the patron's current information need	Planned use of the answer
	Information sources already searched
	Date an answer is needed by
Attributes of the triaging service (the service or triager that received the question)	Number of questions received on a given day
	Number of questions on the same subject that is expected to be received by the service
	Habits that the triager has developed over time in triaging questions
Attributes of the receiving service (the service to which a question is triaged)	Scope of the collection
	Scope of the service: does the service provide general or subject-specific reference?
	Depth of assistance provided: what is provided as or with an answer: citations only, answers, bibliographic instruction, etc.
	Response rate: how many of the questions which the service receives get answered
	Quota for number of questions that can be

Attributes of the receiving service or the answerer	accepted per day
	Subject expertise: areas in which the service or answerer has expertise in answering questions
	Audience served: patrons' affiliation, age, school level, area of residence
	Availability of appropriate information sources that can be consulted in answering the question
	Past performance in quality of answers provided
	Turnaround time for providing an answer
	Country of residence
	State of residence, or other subdivision within the country of residence: e.g., Province in Canada
Attributes of the answerer (the individual expert to whom a question is triaged)	Subject interest beyond subject expertise
	Reference experience
	Customer service expertise
	Expected answer formulation: how the triager anticipates that the answerer will formulate an answer

Some of these attributes are related, and some quite closely. For example, the scope of a service is closely related to the service's subject expertise and the scope of its collection. Naturally a general reference service will need to maintain a collection on a broad range

of subject areas, while a subject-specific service will need to maintain a perhaps more in-depth collection on a narrower range of subjects. Also naturally a reference service will be better able to answer questions within the scope of its subject expertise than outside of it. A reference service's scope, however, is determined prior to either its collection or experience in answering questions. The scope of a reference service is dictated by its mission, and this mission dictates a service's collection development policies and the subject experts that will be employed.

Some other related attributes are the subject of the question and the geographic location of the answerer. One think-aloud respondent's comment about a question concerning the proposed 28th Amendment to the Constitution was:

“... because it's asking about United States government I won't send it to her [a colleague] in Australia – I know she could find it, but someone in the US would have better resources.”

The fact that this question was about the United States government indicated to this respondent that the question should be triaged to an answerer within the United States. Another respondent stated that his service will, whenever possible, forward questions about Texas to digital reference services in Texas, because libraries in Texas have state-wide access to databases about the state, and thus are able to provide more complete answers about Texas than any library outside of Texas would be able to do. Thus again, the subject of a question dictated to the triager the preferred geographic location of an answerer. Thus, there is a great deal of overlap between many of these attributes, and they no doubt influence each other. Nevertheless, they must be treated as separate.

The attributes country and state of residence – which are attributes of both the patron and the answerer – could be considered to be subcategories of a larger attribute, “geographic location.” The triagers, however, treated country and state as important attributes separate from one another, and so, utilizing the constant comparative methodology, these two attributes emerged from the data analysis as separate. For example, one question that was

triaged during one of the think aloud studies was triaged to a digital reference service in Australia, because the patron who submitted the question identified herself as being located in Australia, and the triager stated that the sources that an Australian service would have access to would be different than those in other countries, and more appropriate for an Australian patron, who might want to gain access to those sources herself. On the other hand, there was another question triaged by another service that the triager stated he would ideally triage to the New Mexico state department of education, if only his service had a question-forwarding arrangement with them. Thus, country and state were identified specifically as being attributes that affected the triage process, and not simply geographic location.

The availability of appropriate information sources that can be consulted in answering the question is an attribute both of the receiving service and of the answerer, depending on the type of service for which the answerer is employed. If the answerer is employed by a digital reference service affiliated with a physical library, then the available information sources will be those in the library's collection, and therefore an attribute of the service. On the other hand, if the answerer is employed by an AskA service – that is, unaffiliated with a physical library – then the available information sources will be those that the answerer has access to: their personal collection, the free web, perhaps some fee databases. Thus, the availability of appropriate information sources that can be consulted in answering the question fall under two categories of attributes.

The quota for number of questions that can be accepted per day is also an attribute both of the receiving service and of the answerer. Some services only accept a certain number of questions from other services; the AskERIC service, for example, maintains a list of services to which they may triage questions, and the number of questions those services will accept from AskERIC per day. On the other hand, some experts will only answer a certain number of questions; the VRD service and many of the services affiliated with academic libraries maintain lists of individual experts to whom questions may be triaged, and the number of questions those experts will accept per day.

4.2.2.4. Distribution of Triage Factors

In analyzing the data from the think-aloud protocols, each attribute was counted only once per question. The same attribute was frequently mentioned multiple times by the triager, in the course of triaging a single question. For example, in triaging the question “How do you link an online form to a microsoft access table?,” part of the triager’s verbalization was as follows:

“That is not in our scope. That is going to [another service]. I don’t know that they’re going to do a better job with it than we would here, frankly, but... I mean, I could keep it here, and I could give them resources on Access, or something along those lines.”

The quotes “I don’t know that they’re going to do a better job with it than we would here,” and “I could give them resources on Access, or something along those lines” were both coded as Expected answer formulation, since they both address how the question would likely be answered by two different possible answerers. Thus, the same attribute was mentioned twice in triaging this question, as a way of verbalizing the pros and cons of triaging the question to another service or answering it locally. For another example, in triaging the question “I am looking for a small-group curriculum that will help me work with bullies and/or victims of bullying. I am a middle school counselor,” part of the triager’s verbalization was as follows:

“Middle school counselor. Counseling. We have a clearinghouse on that, so that one goes to them. Any time you see bullying or violence in schools, that kind of thing, self esteem.”

The quotes “bullying,” “violence in schools,” and “self esteem” were all coded as Question subject. The only one of these terms that actually appeared in the question was bullying, but all three terms were used by the triager as a way of verbalizing the scope of the Counseling clearinghouse, to explain why it was appropriate to triage this specific

question there. It does not make sense, however, for a single attribute to affect a triage decision more than once. More than one attribute may affect a triage decision, but a single attribute may affect a triage decision only once. Each attribute was therefore counted only once per question.

The eight attributes of questions were coded as affecting the triage decision 242 times, for the 185 questions triaged during the think-aloud studies: Subject affected the triage decision for 144 questions, Difficulty 46 questions, Specificity 37, Type 6, Multiple questions from one patron 4, Interestingness 3, Prior history of the question 2, and Language 0; this distribution is represented as percentages in Figures 4-3 and 4-4, below. This gives an average of 1.3 attributes that affect triage per question. All thirty-eight attributes were coded as affecting triage 617 times, for an average of 3.3 attributes per question. Thus, more than one attribute influences triagers' decisions for every question triaged, and one of those attributes is always an attribute of the question. This is intuitively reasonable, as it is difficult to imagine how a question could be triaged without actually considering the question itself. Many digital reference services also collect data in addition to the question itself: services that maintain a webform for question submission frequently utilize that webform to collect additional data: geographic location, age or school level, planned use of the information provided, and sources already searched are just some of the data that question submission webforms may solicit. It is intuitively reasonable that the more data is collected along with the question, the more likely that data is to affect triage decisions. Given that twenty-two of the twenty-eight services that participated in this phase of the study (78.6%) maintain question submission webforms, it is reasonable that this additional data affects triage decisions.

The frequencies of these thirty-eight attributes roughly follow a Zipfian distribution, however one slices the data. Table 4-4 lists the number and percentage of questions out of the 185 triaged during the think-aloud studies that were influenced by each attribute. Figure 4-3 shows the distribution of the percentage of the 185 questions collected in phase 1 for which the eight attributes of questions affected the triage decision. Figure 4-4 shows the distribution of the percentage of questions for which all attributes affected the

triage decision. The sum of the percentages corresponding to the attributes in Figure 4-4 is not one, since more than one attribute may affect the triage decision for a single question.

Table 4-4: Attributes that Affect Triage Decisions

Question Attribute	Number of questions	Percentage of questions
Subject of the question	144	77.8
Scope of the service	47	25.4
Difficulty of the question	46	24.9
Generality or specificity of the question	37	20
Expected answer formulation	33	17.8
Availability of appropriate information sources	28	15.1
Quota for number of questions that can be accepted per day	28	15.1
Service: subject expertise	25	13.5
Answerer: subject expertise	25	13.5
Reference experience	23	12.4
Role, job description, or other capacity in which the question is being asked	21	11.4
Turnaround time for providing an answer	21	11.4
Planned use of the answer	18	9.7
Number of questions received on a given day	16	8.6
School level of the patron	15	8.1
Scope of the collection	12	6.5
Customer service expertise	10	5.4
Audience served	7	3.8
Question type	6	3.2

Organizational affiliation	6	3.2
Past performance in quality of answers provided	5	2.7
Multiple questions from one patron	4	2.2
Age of the patron	4	2.2
Response rate	4	2.2
Subject interest beyond subject expertise	4	2.2
Patron: Country	4	2.2
Receiving service or answerer: Country	4	2.2
Interestingness of the question	3	1.6
Number of questions on the same subject that is expected to be received by the service	3	1.6
Prior history of a question with the service	2	1.1
Information sources already searched	2	1.1
Date an answer is needed by	2	1.1
Triager's habits	2	1.1
Depth of assistance	2	1.1
Receiving service or answerer: State	2	1.1
Form in which the patron specifies that the answer should be presented	1	0.5
Patron: State	1	0.5
Language of the question	0	0

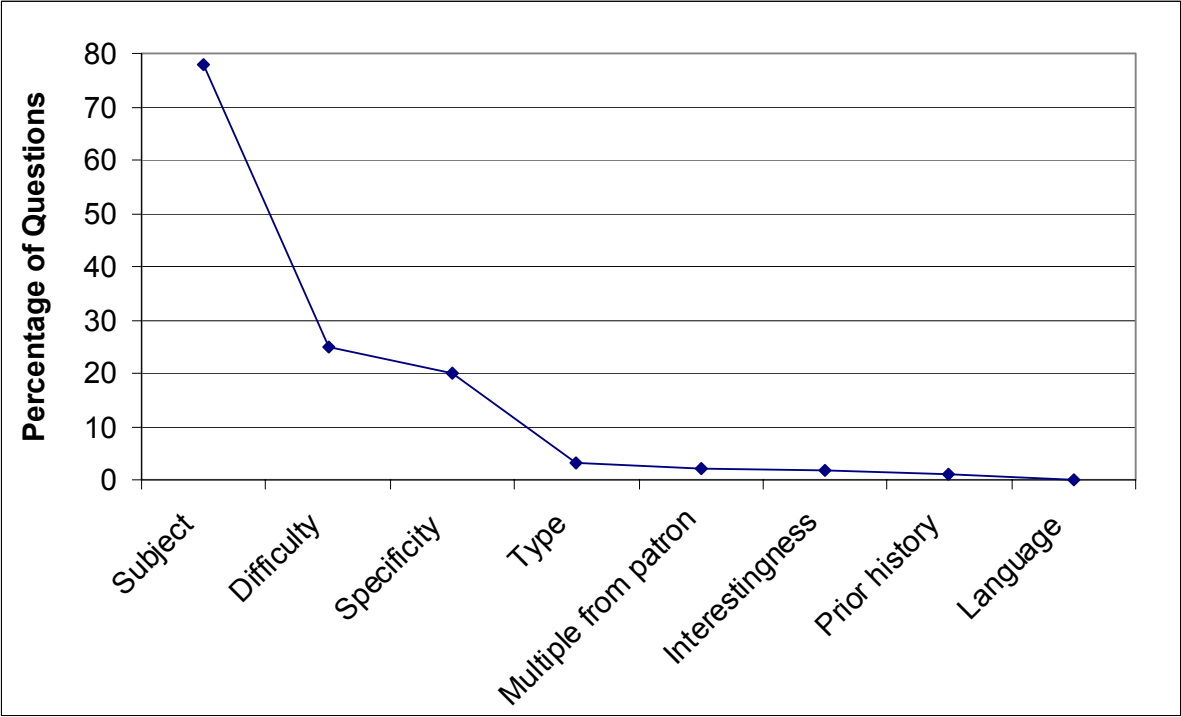


Figure 4-3: Distribution of Question Attributes that Affect Triage Decisions

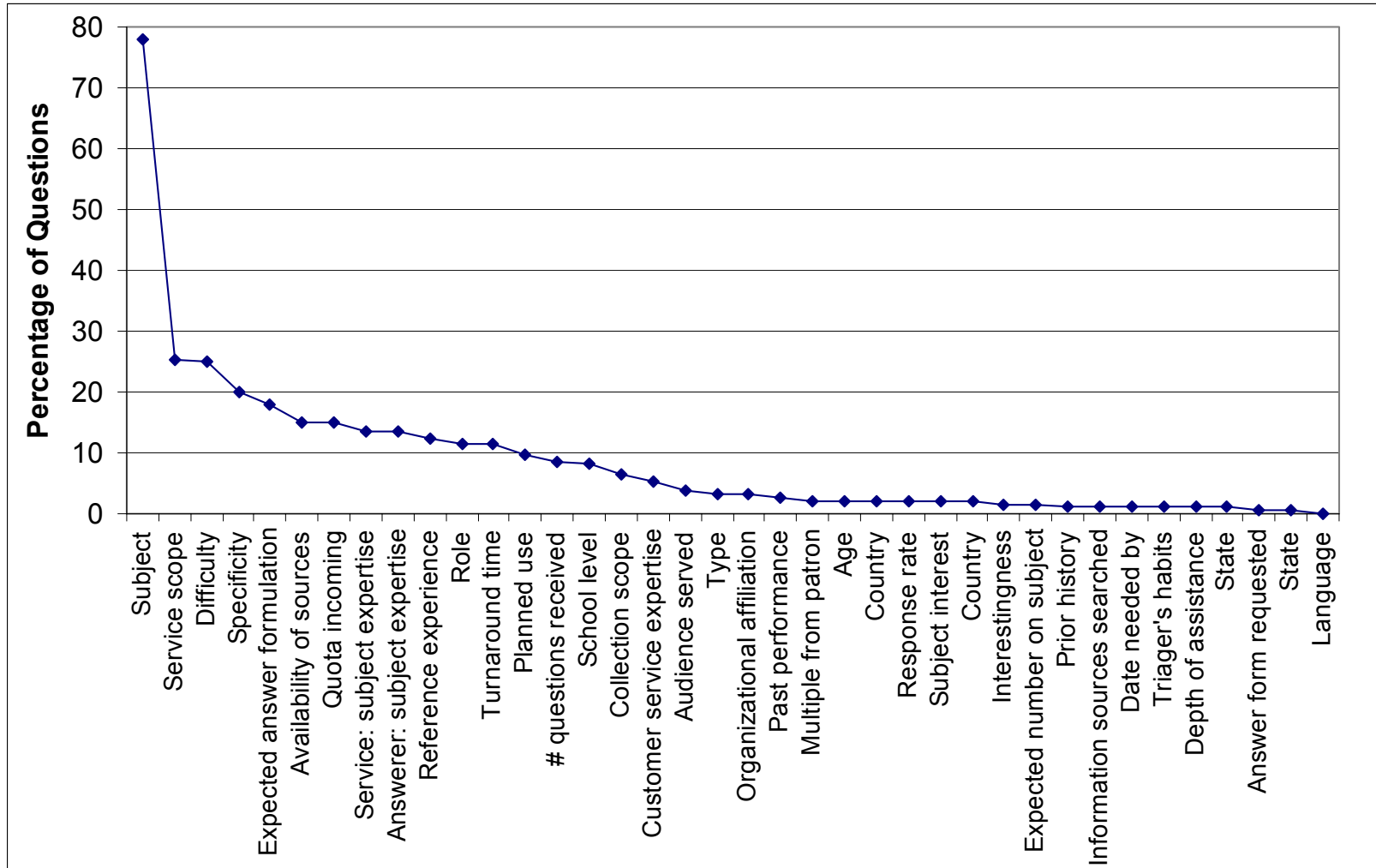


Figure 4-4: Distribution of all Attributes that Affect Triage Decisions

Table 4-5 lists and Figure 4-5 shows the distribution of questions triaged by categories of attributes, across the 617 times that all attributes were coded:

Table 4-5: Questions Triaged by Attribute Category

Attribute Category	Number of questions	Percentage of questions
Question	242	39.2
Receiving service or Answerer	120	19.4
Receiving service	90	14.6
Answerer	70	11.3
Patron	51	8.3
Patron's info need	22	3.6
Triaging service	21	3.4
Answer	1	0.2

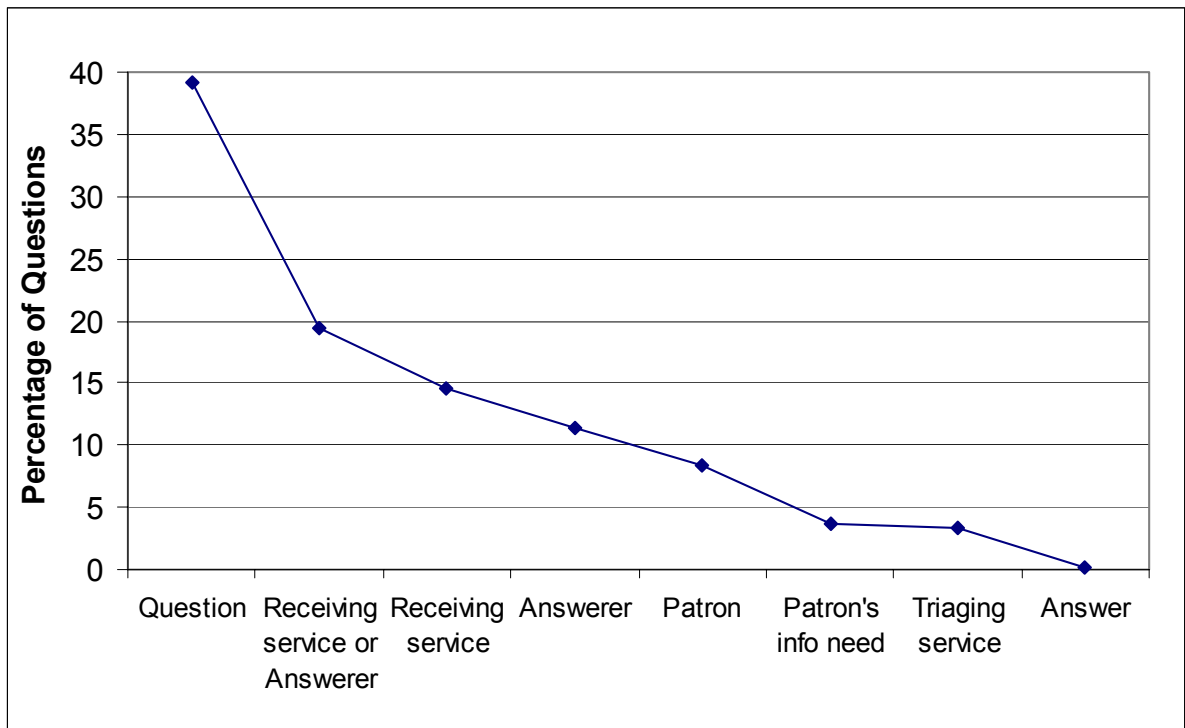


Figure 4-5: Distribution of Questions Triaged by Attribute Category

It was mentioned above that the populations of high-volume and low-volume services may be considered to be two separate populations. An analysis of the eight attributes of questions that affected triage decisions during the think-aloud studies was performed, treating these two populations as separate. The Chi-square test for independence was performed on this data, and showed that at $\alpha = 0.05$, there was not a significant difference between high-volume and low-volume services in the attributes of questions that affect triage decisions. Similarly, an analysis of the eight attributes of questions that affected triage decisions was performed, treating the populations of AskA services and services affiliated with academic, public, and special libraries as separate. The Chi-square test for independence was performed on this data, and showed that at $\alpha = 0.05$, there was a significant difference between types of services in the attributes of questions that affect triage decisions. Further research is indicated with a larger data set of questions from services of all four types, to determine which types of services are significantly different from one another, and if there are significant differences between types of services in other attribute categories that affect triage decisions. Determining the significant differences between the different types of services could allow the development of a future system for automating triage to be customized according to the type of service in which it will be implemented. The possibility of customization in such systems will be discussed in greater detail in chapter 5.

4.2.2.4.1. Comparison with Pomerantz, Nicholson, and Lankes (2003)

The distribution of reasons for making triage decisions shown in Figure 4-4 supports the findings of Pomerantz, Nicholson, and Lankes (2003), who determined fifteen factors that affect the process of routing and assigning reference questions received electronically by digital reference services. These fifteen factors were, in descending order of importance:

1. Subject area of the question
2. The service's area(s) of subject expertise
3. The answerer's area of subject expertise

4. Level / depth of assistance available from the service
5. Number of questions that may be forwarded to the service per unit of time, as set by consortium agreements
6. Response rate of the service
7. The answerer's experience and skill in providing reference service
8. Past performance of the service in providing correct and complete answers
9. The service's turnaround time for answering questions
10. Number of questions that your service may forward to other services per unit of time, as set by consortium agreements
11. Availability of sources to answer the question
12. The answerer's experience and skill with providing customer service
13. Language of the question
14. Scope of the service's collection
15. Question type

Figure 4-6 shows the distribution of all attributes of questions, with Pomerantz, Nicholson, and Lankes' fifteen factors highlighted.

Notice that question subject is the first attribute both in this study's and in Pomerantz, Nicholson, and Lankes' (2003) findings. This indicates that question subject is the single most important attribute that affects triage decisions: not only was it mentioned the most times by triagers during the think-aloud studies, indicating that it is the attribute most attended to in triagers' internal cognitive processes, but triagers consciously recognize that this is the case.

It may be noted in Figure 4-6 that, while question subject is the first attribute in both studies' findings, the remaining fourteen of Pomerantz, Nicholson, and Lankes' (2003) factors are scattered throughout the thirty-eight attributes discovered in this study. Several of the fifteen factors are clustered near the top of the distribution, but the rest are distributed throughout. This indicates that the fifteen factors identified by Pomerantz, Nicholson, and Lankes are not the most important attributes that affect triage decisions,

but rather those of which triagers are consciously aware and able to retrospectively verbalize.

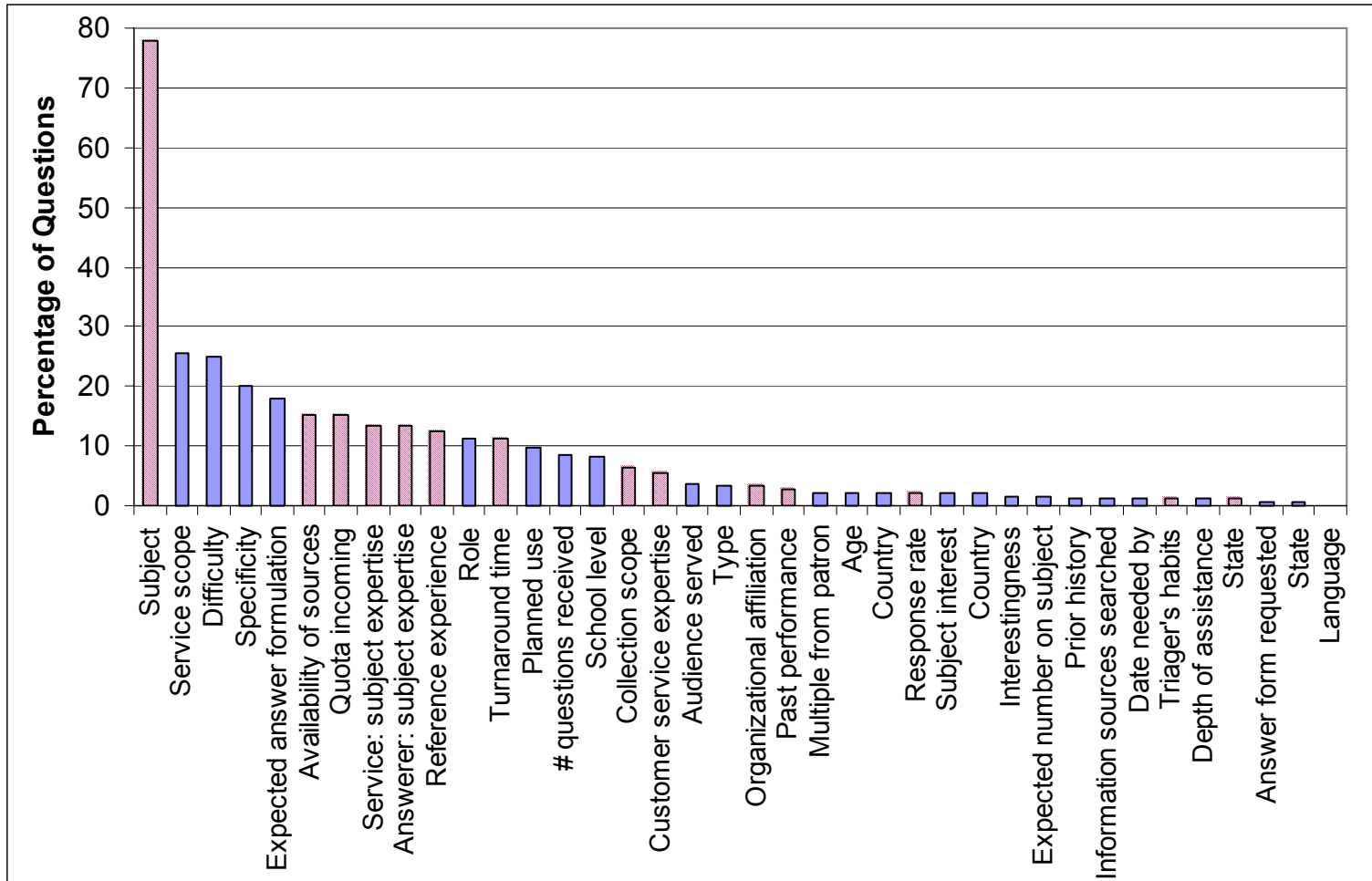


Figure 4-6: Distribution of all Attributes that Affect Triage Decisions, with the 15 Factors from Pomerantz, Nicholson, and Lankes (2003) Highlighted

4.2.2.4.2. Subject

As has already been discussed, this study classified questions according to three taxonomies, and did not classify questions according to a taxonomy of the subjects of questions. This omission is especially glaring in light of the distributions shown in Figures 4-3 and 4-4, where the subject of a question is shown to be the single most important factor that affects triage decisions.

Two reasons that a subject taxonomy was not utilized in this study are that 1) several well-developed classification schemes exist that classify entities by subject, and 2) different digital reference services use different schemes. Thus, either: 1) crosswalks would have had to have been developed between all of the different schemes used by different services, or 2) one scheme would have to have been selected and imposed on all services that did not “natively” use that particular scheme.

A third reason that a subject taxonomy was not utilized in this study is discussed here. The three taxonomies that were utilized to classify questions in this study were three that were identified in the literature from desk and digital reference, question answering, and linguistics, as having been developed specifically to classify questions. Other classification schemes were identified: subject, for example, as well as the types of sources from which the answer may be drawn (see section 2.8.4). Classification schemes according to subject were not originally developed to classify questions, but rather to classify physical materials; it just so happens that they can also be used to classify intellectual entities. Similarly, classification schemes of sources from which answers may be drawn are not designed to classify questions, but rather of documents that may be used in answering questions.

Each of these taxonomies corresponds to a factor that affects triage: for example, sources from which answers may be drawn corresponds to the appropriate information sources that can be consulted in answering the question. Additionally, the taxonomy of functions of expected answers corresponds to the question type attribute, and the taxonomy of

forms of expected answers corresponds to the form in which the patron specifies that he or she would like the answer presented. Indeed, each of the thirty-eight factors that affect triage may correspond to an entire taxonomy, and some of these taxonomies may not yet even have been developed. The possibility of a taxonomy of question difficulty, for example, is discussed in chapter 5. Thus, it may be that an entire taxonomy (existing or as yet undiscovered) may correspond to each of the thirty-eight factors, and this would be a useful avenue for future research towards developing systems to automate the triage process. The three taxonomies according to which questions were classified in this study were, however, those that: 1) had been developed prior to the start of this study, and 2) had been developed specifically to classify questions.

4.2.2.4.3. Language

An unexpected finding is that language did not come up as an attribute according to which questions are triaged even once. The attribute language was one of the fifteen determined by Pomerantz, Nicholson, and Lankes' (2003). Yet, counter-intuitively, it was not elicited even once from the think-aloud studies as affecting the triage process. The researcher believes that this is a case of sampling bias, and not in fact indicative of the importance of the language of a question in affecting the triage process. In order to avoid a language barrier between the researcher and the respondents, the think-aloud studies had to be conducted in English, so respondents were solicited from libraries in nations in which English is spoken. As it turned out, of the 185 questions triaged during the think-aloud studies, 184 of them were written in English, and were triaged by and to individuals who were fluent in English – so naturally the language of the question did not affect the triage process. The 185th question, received by the reference service at the University of Amsterdam Library, was written in Dutch, and was triaged by and to individuals who were fluent in Dutch – though the triager was also fluent in English – so again, the language of the question did not affect the triage process. Sixteen questions were triaged by the National Library of Canada, but all of these were written in English, and as English is one of the languages spoken in Canada, again the language of the question did not affect the triage process. Thus it is the researcher's contention that the

solicitation of respondents from libraries in nations in which English is spoken biased the sampling so that language was cancelled out as an attribute affecting triage. An interesting study could be conducted to test this hypothesis by studying triage in digital reference services for which issues of language are within scope, such as the Ask a Linguist service (linguistlist.org/~ask-ling) and the Slavic Reference Service (www.library.uiuc.edu/spx/srs.htm), or that are affiliated with libraries located in strongly bilingual regions such as southern California and southern Florida, or the Canadian province of Quebec.

4.2.2.5. Origins of Questions

Questions received by digital reference services can come from one of two possible origins: a question can be submitted directly to the service from a patron, or a question can be forwarded to the service from another digital reference service. In the early days of digital reference most services were standalone operations, and the librarians employed by the service answered all of the questions received by the service. Since the late 1990s, digital reference services have begun forming “consortia” to allow services to swap questions. The Virtual Reference Desk (VRD) Network is one such consortium: the VRD’s description of themselves states that:

When a subject specific service receives questions which are out of its stated scope area, it can forward those questions to the VRD Network for assistance. If a question cannot be addressed by another participating service, it will be handled by one of the VRD Network Information Specialists. (www.vrd.org/network.shtml)

There are a number of other such consortia, both national and local: The QuestionPoint collaborative reference is managed by the Library of Congress, while the Metropolitan Cooperative Library System (MCLS) is an association of libraries in the greater Los Angeles area. Of course, the specific services that are allowed to swap questions within these consortia are only those services that participate – as of this writing VRD has

fifteen participants (Bennett, personal communication, 2002), MCLS has 32 full and 24 associate members (www.mcls.org/nonmembers/aboutmcls/), and “over 300 libraries are using QuestionPoint, including users in Australia, Canada, China, England, Germany, the Netherlands, Norway and Scotland” (Penka, 2003, QuestionPoint section, ¶ 3). Within these consortial services, then, questions will routinely be forwarded between participating members. Even digital reference services that do not officially participate in such consortia may sometimes forward questions to other services: one think-aloud respondent stated that even though her service is not officially a participant in any digital reference consortium, she will triage questions to the Internet Public Library, the New York Public Library (her institution is located in New York City), and even the library at the patron’s own institution, if the patron is affiliated with another institution and she deems it appropriate.

The break-down of the origins of questions received by all services studied in the think-aloud studies is as follows:

- Patron: 91.9%
- Other service: 8.1%

4.2.2.6. Destinations of Questions

The other side of the triage equation, the destination to which a question is triaged, is of course the primary focus of this study. Different digital reference services studied in this phase triaged questions to different sets of recipients. The VRD triaged questions to its network of fifteen participating AskA services as well as to volunteer reference and subject experts, so that the “triagee” may be a service or a specific individual. The AskERIC service triaged questions to its sixteen subject-specific clearinghouses, several adjunct clearinghouses and support components, and to specific individuals employed by “AskERIC central” (the Clearinghouse on Information and Technology), where the triage process is performed, as well as occasionally to other AskA services. Digital reference services affiliated with academic libraries triaged questions primarily to specific

departments within the library or specific libraries on campus, and only occasionally to other AskA services.

Because of the wide range of recipients to which different services triage questions, it was difficult to make any generalizations about these recipients or their characteristics, except to determine whether the recipient is internal or external to the digital reference service itself or the organization with which the service is affiliated. Lankes (1998a) refers to triage between services “meta-triage” (p. 166). That term will not be used in this study, however, in order to retain the equivalence between triage where the recipient is internal or external to the service. That equivalence is important for two reasons: first, the outward action of triaging a question is nearly identical whether the recipient is internal or external to the service; the term “outward action” is used here to refer to the physical action performed by the triager, as distinct from the triagers’ internal cognitive processes. In an email-based digital reference service, as all of the respondents in this phase were, forwarding a question is technically simple; every email application has a Forward button and an addressbook, and they work the same regardless of whether the recipient is within the building or on the other side of the world. Second, like consortia of physical libraries, consortia of digital reference services (in which services may forward questions to one another) are entities at a level of analysis larger than that of the individual service. Thus, at the consortial level of analysis, triaging questions between services can be seen as triaging questions within a single large service, the consortium.

According to this simple categorization of triage recipients, questions received by digital reference services can be triaged to one of two possible destinations: within the service, or to another digital reference service. The break-down of the triage destinations of questions from all services studied in the think-aloud studies is as follows:

- To an answerer within the service: 81.1%
- To another service: 18.9%

4.2.2.6.1. Destinations of Questions According to the Volume of the Service

The break-down of the triage destination of questions from high- and low-volume services is presented in Table 4-6. This table should be read down, not across. The columns sum to 100%, the rows do not.

Table 4-6: Break-down of Triage Destinations of Questions from High- and Low-Volume Services

	High-volume services	Low-volume services
To an answerer within the service	76.4%	92.7%
To another service	23.6%	7.3%

According to Table 4-6, both high- and low-volume digital reference services triage a greater percentage of questions internally than externally. This is easily understood, as a reference service would not be much of a reference service if it answered fewer questions than it forwarded to other services. An exception to this would be a question-swapping consortium, such as the VRD or QuestionPoint services. QuestionPoint, however, is not strictly speaking a digital reference service, as it does not answer any questions itself; it is strictly a triage center for routing and assigning questions to other participating digital reference services. The VRD, on the other hand, is both a triage center and a digital reference service, and it triaged nearly twice as many questions internally as externally during this study.

The numbers in Table 4-6 indicate that the ratio of questions triaged internally versus externally by low-volume services is approximately 10:1, while the ratio of questions triaged internally versus externally by high -volume services is approximately 3:1. This too is easily understood. As the number of questions received by a digital reference service increases, the number of questions triaged out of the service is also likely to increase, though it does not necessarily follow that the *percentage* of questions triaged

out will increase. As the number of questions received by a digital reference service increases, however, the service has greater motivation to join question-swapping consortia, since, as in physical libraries, hiring more staff to answer the greater number of questions is often prohibitively expensive, and such consortia are one means for services to handle an increasingly large number of questions without increasing the size of the service. Once a service has joined a question-swapping consortium, it then has increased options for triaging questions out to other services. As with any new technology, it may take the service some time to figure out how to utilize these options, to determine what types of questions should be triaged out. Over time, however, it seems likely that services that are members of question-swapping consortia will learn to utilize them, and will triage out a greater percentage of questions than they did before. This hypothesis could be tested by studying a digital reference service prior to and after joining a question-swapping consortium. It seems logical, however, that participating services in question-swapping consortia – of which there are a greater number among the high-volume services studied in this study – triage a greater percentage of questions externally than those that do not participate in such consortia.

4.2.2.6.2. Destinations of Questions According to the Type of the Service

The break-down of the triage destination of questions from digital reference services affiliated with different types of organizations is presented in Table 4-7. This table, like Table 4-6, should be read down, not across. The columns sum to 100%, the rows do not.

Table 4-7: Break-down of Triage Destinations of Questions from High- and Low-Volume Services

	Academic libraries	AskA services	Public libraries	Special libraries
To an answerer within the service	82.8%	85.4%	100%	58.8%
To another service	17.2%	14.6%	0%	41.2%

According to Table 4-7, digital reference services affiliated with all types of libraries (or, in the case of AskA services, unaffiliated with any library) triage a greater percentage of questions internally than externally. Again, this is easily understood, as a reference service would not be much of a reference service if it answered fewer questions than it forwarded to other services.

The most noticeable thing about Table 4-7 is the range of ratios of questions triaged internally to externally: the ratio of questions triaged internally versus externally by services affiliated with academic libraries is approximately 5:1, AskA services are approximately 6:1, public libraries are 1:0, and special libraries are approximately 1.5:1. The fact that there is a range in the ratios of questions triaged internally to externally is to be expected, as different types of libraries have different expectations for reference service: it is to be expected, for example, that AskA services, many of which participate in question-swapping consortia, would triage a greater percentage of questions externally than would public libraries, many of which do not participate in such consortia. What is surprising is that AskA services did not have the greatest percentage of questions triaged externally; given the fact that AskA services exist entirely online, it might have been expected that they would utilize their networked nature to promote inter-service collaboration more than any other type of digital reference service. Particularly striking is the nearly equal ratio of questions triaged internally and externally by services affiliated with special libraries. This is clarified when one considers the identity of the special

libraries who participated in this study. These services cannot be named, since the researcher promised anonymity to all participating services. That said, however, these special libraries would be more accurately identified as government libraries (in the United States and other nations). Thus, the nearly equal ratio of internally and externally triaged questions is made clear, since, if there is an arena in which partnerships between organizations is even more important than in library work, it is government.

4.2.2.6.3. Attributes that Affect the Destinations of Questions

Question triage is, to a certain extent, analogous to interlibrary loan (ILL), in that objects (questions or books) are moved between services. There are some obvious differences between triage and ILL, particularly in that in ILL the recipient instigates the request, whereas in triage the sender instigates. But let us ignore these differences for the moment and focus on the fact that both triage and ILL may be treated as networks of nodes. In any network, there are two possible levels of analysis: the individual node, and the network as a whole. In a triage network (a consortium), there is a “conservation of questions” that must occur, just as in an ILL consortium there is “conservation of books”: all questions or books that are sent by one node (service) must be received by another service. Thus, over the entire consortium, the number of questions or books remains constant. On the other hand, at the individual service level, the number of questions or books does not need to remain constant: some services may triage out more questions than they receive, just as some libraries may receive more ILL requests than they make. All of this is to explain why the percentage of questions triaged to other services is so different than the percentage of questions received from other services. The level of analysis for the origins and destinations of questions in this study was the individual service, so it is not necessary that these numbers be equivalent. Further, the services that participated in this study are not all members of one single triage consortium, which is all the more reason that these numbers need not be equivalent, even across all services.

Figures 4-7A, 4-7B, and 4-7C show the distribution of attributes of questions that affect triagers’ decisions, between services, within services, and in total. As these figures show,

not only do the same attributes of questions affect triagers' decisions to triage a question to an answerer within their own service or to another service, but the distribution of these attributes is consistent between these groups. This is particularly noticeable in comparing the Within services and Total distributions, which are nearly identical. The Between services and Total distributions are also quite similar, though the attributes Prior history and Question type do not appear in the Between services distribution. It makes sense that the prior history of a question would not be a factor in triaging that question between services: unless the question has been triaged multiple times, the prior history is with the triaging service specifically. This seems to indicate that when a question has a prior history with a service, it is generally answered by that service, and not triaged to another service. It is less clear why the question type does not appear here as a factor in triaging questions between services. It is possible that triagers simply do not recognize or know how to articulate that their decision process is taking a question's type into account, since certain question types in fact are triaged consistently (see section 4.4.5).

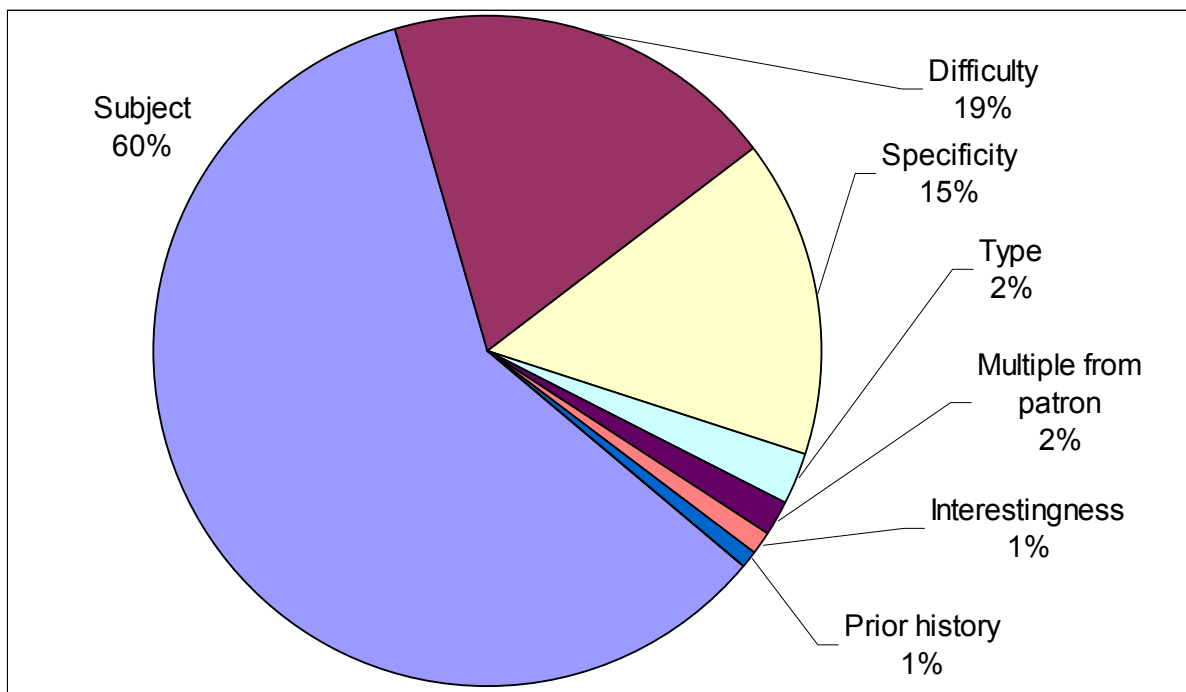


Figure 4-7A: Distribution of Question Attributes that Affect Triage Decisions, in total

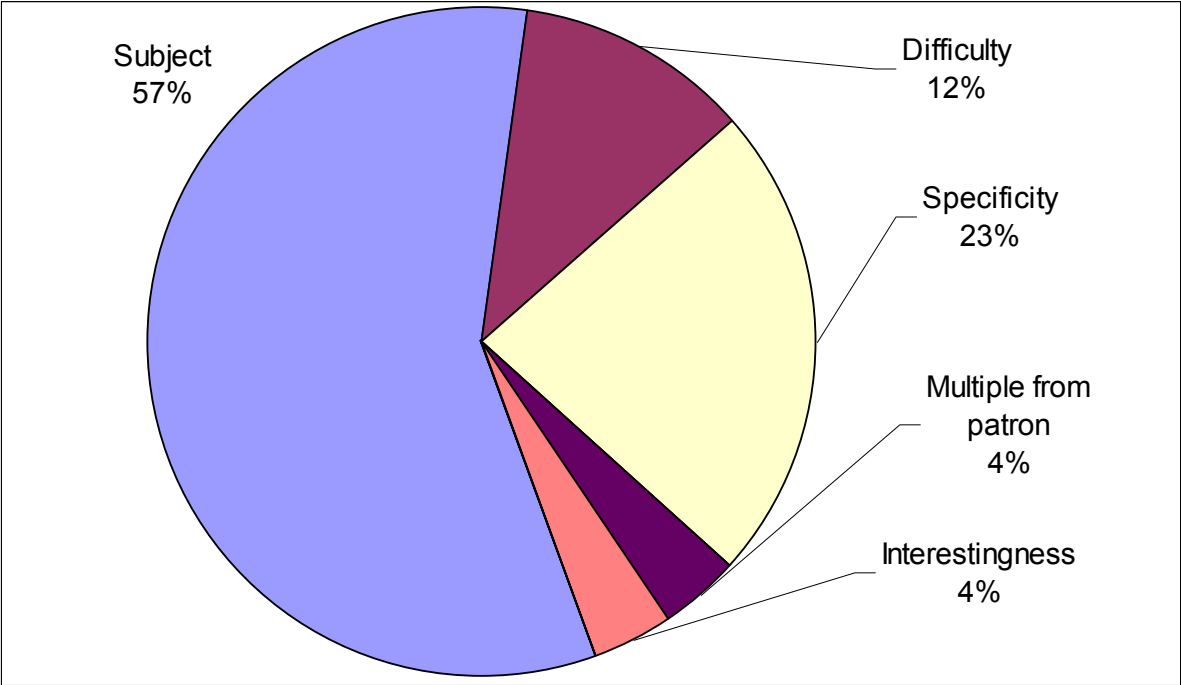


Figure 4-7B: Distribution of Question Attributes that Affect Triage Decisions, Between Services

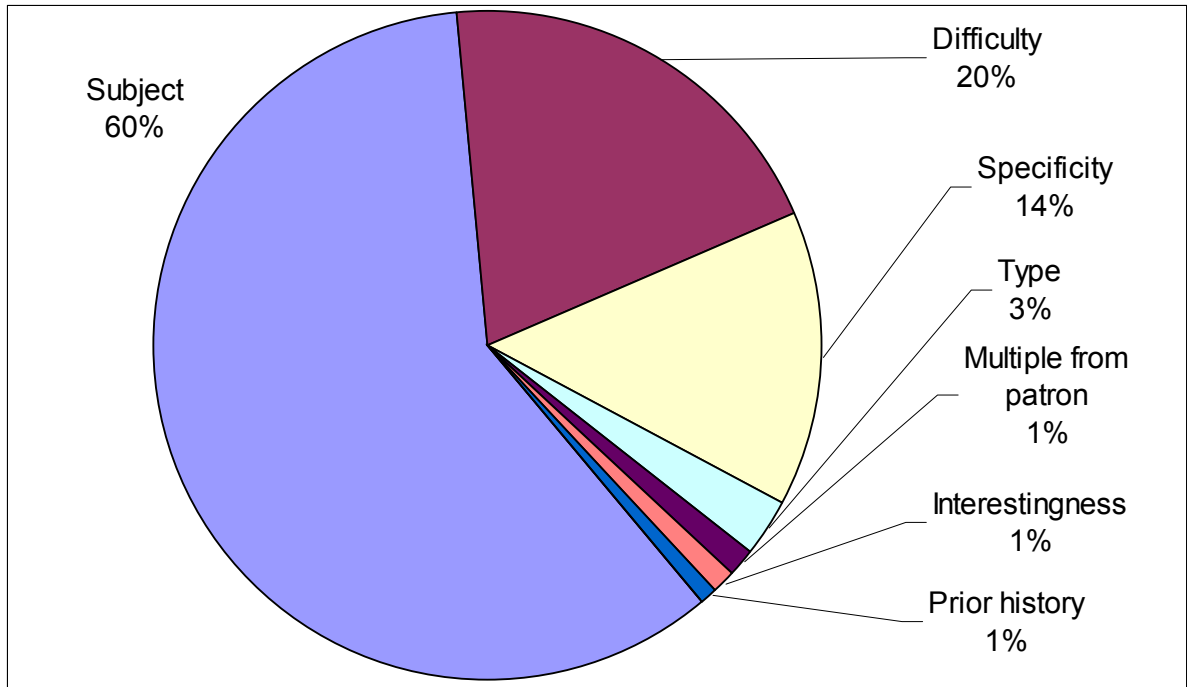


Figure 4-7C: Distribution of Question Attributes that Affect Triage Decisions, Within Services

4.2.3. Post-Think-aloud Survey

The attributes of the twenty-eight digital reference services that participated in the think-aloud studies were discussed above, in section 4.2.1. This breakdown was presented by country and by type of library with which the service was affiliated. Pomerantz and others (forthcoming), however, make a case for a more sophisticated analysis of digital reference services, based upon functional, rather than organizational, characteristics: specifically, the services' use of automation in the process of providing asynchronous digital reference.

The survey that was administered after the think-aloud surveys collected data about seven of the nine key characteristics that Pomerantz and others (forthcoming) use to differentiate three different types of services. These seven characteristics are the bottom

seven listed in Table 4-8. The other two characteristics, shaded at the top of Table 4-8, were determined before a think-aloud study was conducted with any given service. These nine characteristics are listed in Table 4-8 for the twenty-eight services that participated in the think-aloud studies.

Table 4-8: Post-Think-aloud Survey Results

Process	Percentage of services that perform process
Maintains a webform for question submission.	78.6
Automatically sorts questions to experts.	0
Verifies email addresses prior to working on a response.	14.3
Automatically generates an acknowledgement that the question was received.	75.0
Has the ability to detect follow-up questions.	3.6
Automatically searches a knowledge base when a question is received.	0
Stores question-answer sets in a knowledge base.	57.1
Patrons can pick up their responses on the web.	3.6
Automatically tracks the progress or state of a question.	0

The large percentage of services that maintain a webform for question submission is consistent with Pomerantz and others' (forthcoming) finding that 83% of services surveyed maintain a question submission web form. This study's findings are also consistent with those of Pomerantz and others, in terms of the large percentage of services that automatically generate an acknowledgement that the question was received: Pomerantz and others report that 94% of services that maintain a webform for question submission generate an acknowledgement via either email or a webpage.

The finding that no services that participated in this study automatically sort questions to experts of course reflects the criteria that was utilized in soliciting services for participation; specifically, that the service performed triage manually, not automatically. Additionally, no services automatically search a knowledge base when a question is received, or automatically track the progress or state of a question. This is an indication that the overall composition of services that participated in this study are what Pomerantz and others' (forthcoming) refer to as "Low Tech/High Touch" services: those services that rely heavily on human intermediation to handle questions throughout the entire digital reference process. Among Pomerantz and others' respondents, this group was composed primarily of academic and public libraries. Academic and public libraries together make up 67.8% of the services that participated in this study; so again, this study's findings are consistent with Pomerantz and others

4.2.4. Zooming In on the Triage Process

Within the past few years, a few models of digital reference have been proposed and studied (Robinson, 1990; Lankes, 1998a; McClennen, 2001; Pomerantz et al. (forthcoming)). The current drawback of these models is that they offer a view from a very high altitude, as it were, encompassing a broad domain in little detail. Pomerantz, Nicholson, and Lankes (2003) began the task of delving into one specific process described in these models, the triage process. The present study has further decomposed the triage process, by determining the full range of attributes of questions and other factors that contextualize the question. Using these attributes, a detailed model of the triage process is proposed here, consisting of the rules for actions taken on questions received by digital reference services. Thus, this study builds on and develops current models of the digital reference process by "zooming in" on the triage process.

Several different detailed models of the triage process as performed in specific services and types of services were developed. For example, a detailed model of the triage process was developed for the AskERIC service specifically, since the triage process in that

service is unique, and sufficiently unlike the triage process as it is performed in any other service. A detailed model of the triage process was also developed for all services affiliated with academic libraries, as the triage process in all academic libraries' services was sufficiently similar that a general academic library model could be developed. There is thus not one model of the triage process, but several. All of these models, however, have similarities, due to the fact that there are only thirty-eight possible decision points. Thus, it was possible to create a general model of the triage process, valid for all services surveyed in this study. This model is shown in Figure 4-8.

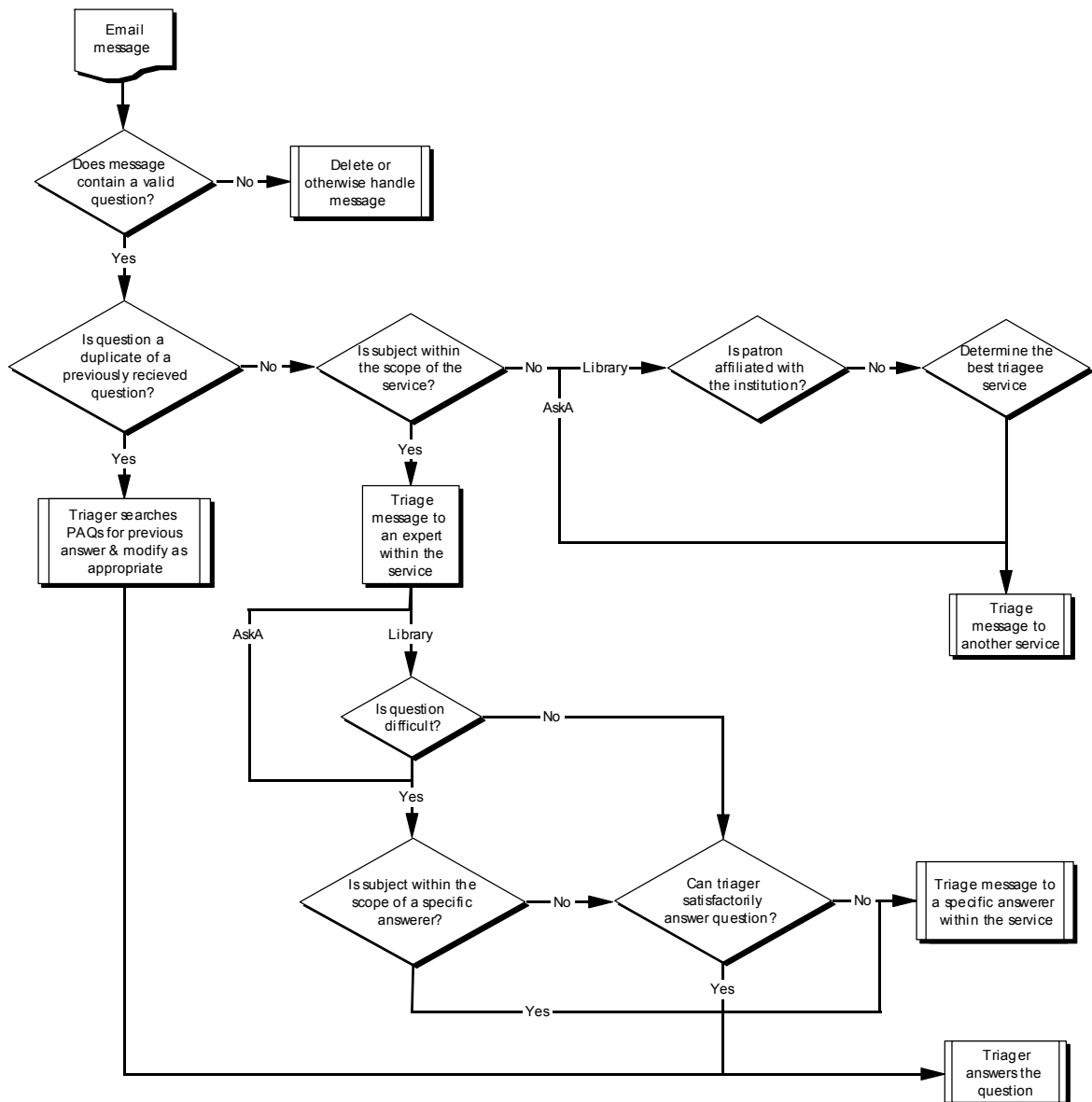


Figure 4-8: General Model of the Triage Process

This general model of the triage process begins with the receipt of a question by a digital reference service. The first step is that all non-questions must be filtered out: such non-questions include viruses, advertisements, server error messages, spam, “thank you” messages from patrons, and a variety of other types. This is a necessary first step in providing digital reference service, but unknown in question answering (QA) systems. This is a necessary step in digital reference service because, as discussed in section 3.3.2,

approximately 15% of the total emails received by digital reference services are non-questions, and it would be a waste of human intermediaries' time and effort to spend time responding to non-questions. This is an unknown step in QA systems because most such systems have not been implemented outside of the laboratory of the TREC QA track, and it may safely be assumed that all queries provided to TREC QA systems are valid. Further, designers of QA systems have inherited the assumption made by designers of information retrieval (IR) systems, that if a question or a query cannot be answered by the system, that it is an acceptable outcome for the system to provide no documents or passages to the user. In any future automated triage or digital reference question answering system, it would *not* be an acceptable outcome for the system to not triage or not answer the question. Thus, just as it will be necessary for automated triage systems to determine whether a question is within the scope of a service, so it will be necessary for such systems to determine whether a putative question is in fact valid.

This general model of the triage process is of course highly simplified, as evidenced by the fact that not all thirty-eight factors that affect triage decisions are represented. In order to create a general model of the triage process, valid for all services surveyed in this study, many steps had to be collapsed into one, or skipped entirely. For example, the decision point *Determine the best triagee service* (in the top right corner of Figure 4-8) is itself a complex process, which may explode into a number of factors including: the specific services to which the receiving service may triage questions, and all of the attributes of the receiving service. For another example, the decision point *Is subject within the scope of a specific answerer?* (near the middle at the bottom) may explode into some or all of the attributes of the answerer. Which specific factors affect triage decisions is of course specific to the service in which triage is being performed.

Additionally, this model makes a distinction at two points between actions taken by AskA services and actions taken by services affiliated with libraries. One such split is at the decision point *Is subject within the scope of the service?* If an out-of-scope question is received by an AskA service, it is forwarded to another service for which it is in scope and with which the receiving service has a question-swapping agreement. On the other

hand, if an out-of-scope question is received by a service affiliated with a library, it may still be answered by that service, if the patron is affiliated with the organization in which the library exists (an academic institution, a corporation, etc.) – particularly if the patron is an important figure in the organization, such as a dean or a CEO. Another split between actions taken by AskA services and services affiliated with libraries is at the decision point *Triage message to an expert within the service*. In services affiliated with libraries, if the question is not difficult and the triager can satisfactorily answer the question, the triager will generally answer it, often interrupting the triage process to do so. On the other hand, while triagers may also be answerers in AskA services, and a triager may triage a question to him- or herself, there is a split between the time in which one individual is performing triage and the time in which that same individual is answering questions.

One of the goals of this study was to draw up a set of rules for the performance of triage, which could be utilized as the basis for designing and building a system to automate part or all of the triage process. This model, and the more detailed models of the triage process as performed in specific services and types of services, may be utilized as the basis for designing systems to automate the triage process. As was discovered in the think-aloud studies, every digital reference service performs triage slightly differently, so no algorithm based on these models could be utilized in any service – other than the ones that participated in the think-aloud studies – without customizing the model to service-specific idiosyncrasies in the performance of triage. These models, however, may serve as the basis for “template” triage algorithms that may be customized according to service-specific rules.

4.3. Phase 2: Identification of Question Taxonomies

This section discusses the identification of four taxonomies of questions through an extensive review of the literature from several fields that deal with questions, and the findings from Phase 2 of the study.

4.3.1. Classification of Questions

This section describes the findings from the classification of the data set of questions sampled from the Virtual Reference Desk (VRD) archives. This classification was performed according to three of the four taxonomies of questions identified in the literature, as they existed in the literature – that is, before being modified. The pre-modified versions of these taxonomies were not identical to the form of these taxonomies subsequently utilized in Phase 3 of this study. This phase of the study was conducted in order to classify questions received by a digital reference service according to the three taxonomies of questions identified in the literature, as a methodology for evaluating these taxonomies for their appropriateness for this task, and, if these taxonomies were found to be appropriate, which they were, to modify them to clarify their scope notes.

4.3.1.1. Taxonomy of Wh- Words

The taxonomy of wh- words, simple though it is, is remarkably expressive. Of the 396 questions in the data set, 83.3% could be classified according to this taxonomy. In other words, 83.3% of the questions in the data set were phrased as questions, utilizing a wh-word in the question. Table 4-9 lists and Figure 4-9 shows the distribution of the percentage of questions classified according to each class in this taxonomy.

On the other hand, 16.7% of questions could *not* be classified according to this taxonomy: the group of unclassifiable questions therefore contained the fourth-largest percentage of questions according to this taxonomy, and more than three times as many questions as the groups of unclassifiable questions according to the taxonomies of functions and forms of expected answers.

Table 4-9: Questions Classified According to the Taxonomy of Wh- Words

Question Classes	Percentage of Questions Classified
What	28.3
How	20.7
Where	18.2
Who	5.3
Which	4.8
Why	4.3
When	1.8

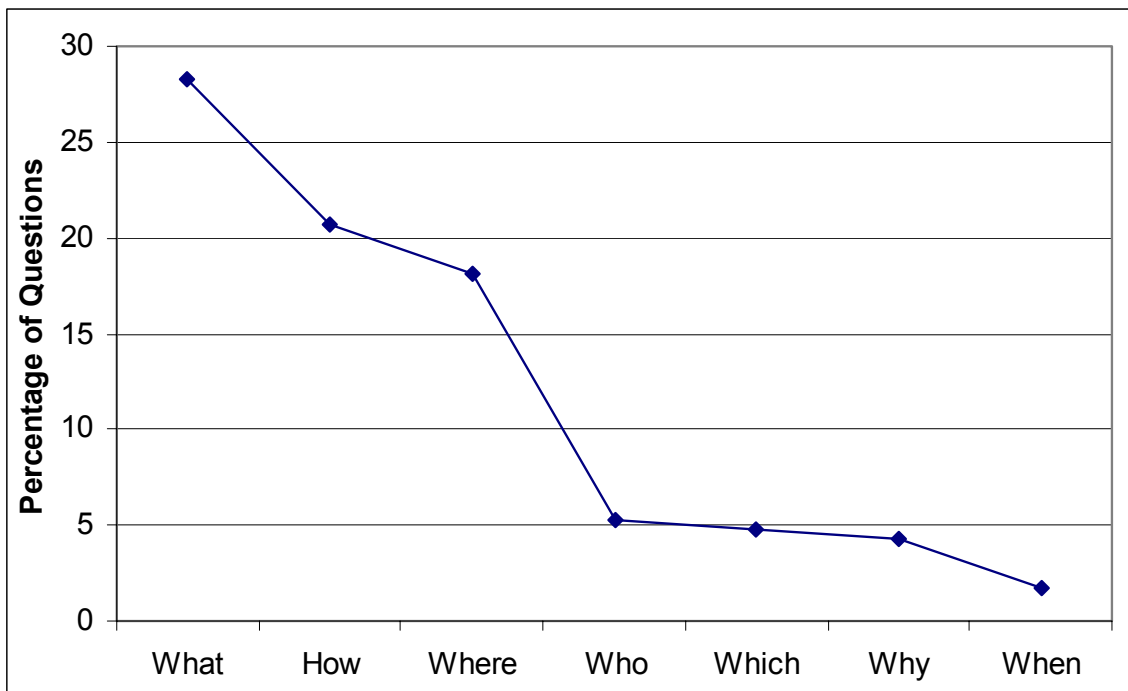


Figure 4-9: Distribution of Questions Classified According to the Taxonomy of Wh- Words

Many of the unclassifiable questions according to this taxonomy are phrased not as questions at all, but rather as statements of the forms: “Can you help me find ...” or “I’m

looking for information on ...” Statement of these forms are similar to the Request/Directive and Assertion types from the taxonomy of functions of expected answers. Indeed, of the unclassifiable questions, 30.3% were classified either as Request/Directive and 18.2% as Assertion according to the taxonomy of functions of expected answers (for a total of 48.5%). These are common question forms in reference (Dewdney and Michell, 1996), yet they cannot be expressed according to the taxonomy of wh- words, for the simple reason that they do not contain a wh- word, and cannot be rephrased in such a way that they will contain one. English speakers tend to think of the five Ws as encompassing all questions that it is possible to form in English, and yet the study of discourse analysis tells us that an utterance that has the force of a question is not necessarily phrased as a question. Such questions will therefore fall between the cracks of the taxonomy of wh- questions. Examples of such questions are: “Tell me information about socretes [*sic*] an astrologer,” and “I would like to know more about the culture of the mexican women from mexico.”

A noticeable divide occurs in the distribution of questions according to the taxonomy of wh- words: What, How, and Where questions each account 18% or more of the questions classified, while Who, Which, Why, and When questions each account for 5% or less. The predominance of What questions may be explained by the fact that the What class may be decomposed into three sub-categories: What-selection, What-description, and What-quantity. These three sub-categories were “unpacked” from the What class during the evaluation of the taxonomies, and utilized in the classification task in Phase 3 of this study.

4.3.1.1.1. Intersection of the Taxonomies of Wh- Words and Functions of Expected Answers

The large percentage of How questions may be explained by the fact that in English, some questions are phrased using the word “How,” when in fact they may be more appropriately classified as What questions; for example, questions of the form “How many” and “How much.” Thus, a significant percentage of the questions classified as

How questions may be more appropriately reclassified as What-quantity questions. Indeed, 39% of How questions were also classified as Quantification questions according to the taxonomy of functions of expected answers. Some questions from the data set classified as How that may be more appropriately reclassified as What-quantity questions are: “How many counties are in the United States?” and “How many different kinds of cows are there in the world?”

4.3.1.1.2. Intersection of the Taxonomies of Wh- Words and Forms of Expected Answers

An unexpected finding of this analysis was the large percentage of questions classified as Where questions that were also classified also as Readers advisory (61.1%) and Directional (18.1%) questions according to the taxonomy of forms of expected answers. Questions at this intersection are those asking for assistance in locating information sources, such as the following: “Can you direct me to Internet resources concerning Benny Benson?” and “I am trying to locate a 2000 calendar on particle physics.” What is surprising about the classification of these questions is not their form, but how many there are. Librarians at physical reference desks are frequently asked for directions – the classic case being the ubiquitous “where’s the bathroom” question. It is surprising, however, that directional questions are equally common on the Internet (except for the bathroom question), where there is no physical space within which the reference expert can direct the patron. This seems to lend support for the argument that has been made in the Information Architecture community for years that users perceive online environments as virtual spaces (Fleming, 1998; Rosenfeld and Morville, 1998). An alternative explanation is that users perceive digital reference experts similarly to the way in which reference librarians at a physical reference desk are perceived: patrons may ask Readers advisory and Directional questions to desk reference librarians because, by dint of being a reference librarian specifically, the patron may assume that the librarian has an understanding and mastery of the information sources in the collection, and will be able to direct the patron appropriately. If digital reference experts are perceived similarly by

digital reference users as having mastery of electronic information sources, then it is appropriate that users would similarly ask Readers advisory and Directional questions.

4.3.1.2. Taxonomy of Functions of Expected Answers

As can be seen in Figure 4-10, the distribution of questions according to the taxonomy of functions of expected answers follows a Zipfian distribution (Zipf, 1949).

Table 4-10: Questions Classified According to the Taxonomy of Functions of Expected Answers

Question Classes	Percentage of Questions Classified
Concept completion	31.6
Request/directive	17.4
Quantification	11.1
Verification	9.6
Assertion	7.8
Instrumental/procedural	5.8
Interpretation	2.3
Enablement	1.8
Causal consequence	1.5
Disjunctive	1.3
Feature specification	1.0
Comparison	1.0
Expectational	1.0
Definition	0.8
Goal orientation	0.8
Example	0.5
Causal antecedent	0
Judgmental	0

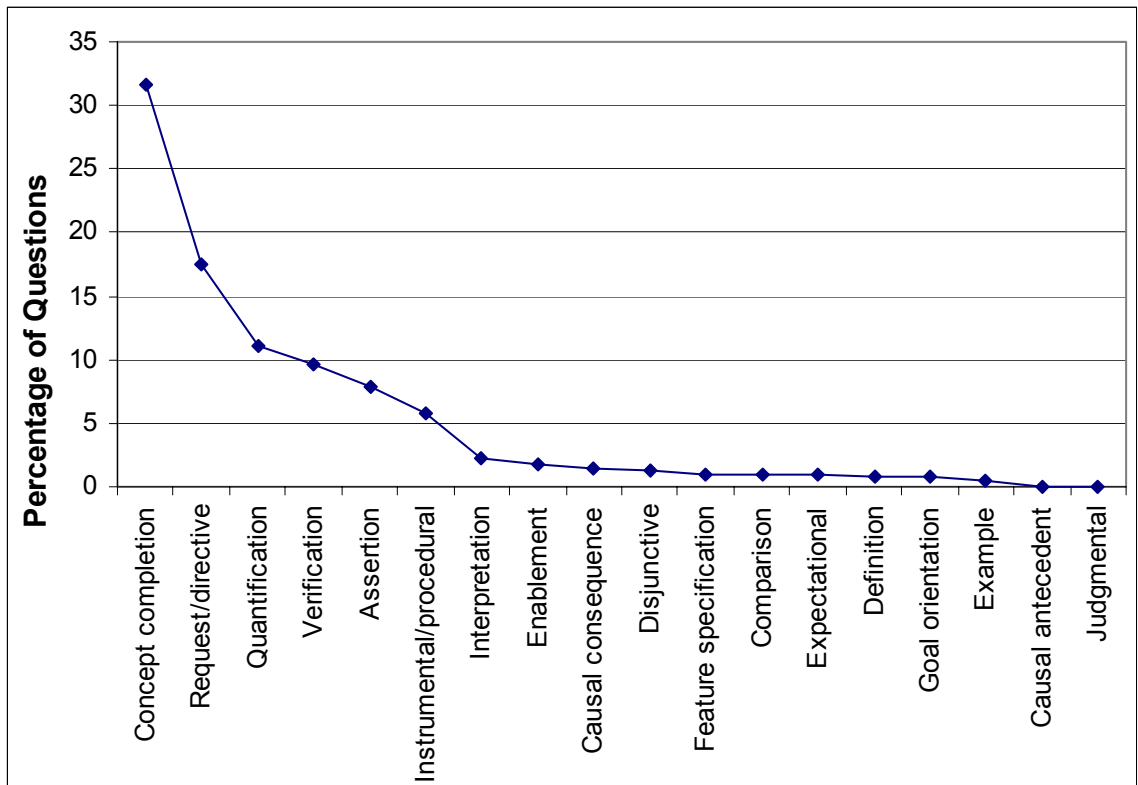


Figure 4-10: Distribution of Questions Classified According to the Taxonomy of Functions of Expected Answers

The class containing the greatest percentage of questions was Concept completion, at 31.6% of the questions classified. The Concept completion class is a “fill-in-the-blank” type of class, and in fact for the sake of clarity this class was renamed as such for Phase 3 of this study. The scope note for this class states that it requires the determination of the subject of a sentence, though perhaps it would be more accurate to say that this class requires the determination of the *focus* of a question; the focus of a question, according to Chen and others (2001) is “the type of answer expected” (p. 481). Indeed, the Concept completion class requires the determination of simple and clear foci of questions: people, places, or other named entities, or times and dates. Graesser, Person, and Huber (1992) describe this class as containing questions of the form: “Who? What? When? Where?” (p.

172). And indeed, every Concept completion question was classed according to the taxonomy of wh- words as Who, What, When, Where, or Which.

4.3.1.2.1. Intersection of the Taxonomies of Functions of Expected Answers and Wh- Words

The Request/Directive class contained 17.4% of the questions, making it the second largest class – though containing only slightly more than half the number of questions of the largest class. As discussed in the previous section, there is a great deal of overlap between the Request/Directive class and unclassifiable questions in the taxonomy of wh- words: 29% of Request/Directive questions were unclassifiable according to the taxonomy of wh- words. More interesting is the fact that 39% of Request/Directive questions were classified as Where questions according to the taxonomy of wh- words. Some examples of these types of questions are: “Can you direct me to Internet resources concerning Benny Benson?” and “Could you please try to direct me appropriate sources [about radar guns].” Questions of this form would be considered Directional if asked at a library reference desk, as they ask for the librarian to direct the patron to materials. The same is true online, only the materials are not in a library’s collection, but on the Internet.

Another interesting point of intersection between this taxonomy and the taxonomy of wh- words is the Instrumental/procedural class. The scope note for this class states that it encompasses questions of the following form: “What instrument or plan allows an agent to accomplish a goal?” It therefore makes a great deal of sense that the majority (91.3%) of questions classified as Instrumental/procedural questions would be classified as How questions according to the taxonomy of wh- words, such as the following: “How do I contact William J. Henderson, Postmaster General of the US?” and “How does a seismograph sense vibrations of an earthquake before an earthquake surfaces?”

4.3.1.3. Taxonomy of Forms of Expected Answers

Similar to the distribution according to the taxonomy of functions of expected answers, the distribution of questions according to the taxonomy of forms of expected answers appears to follow a Zipfian distribution. With fewer classes in this taxonomy, it is slightly more difficult to see the curve of the distribution – however, even in the distribution of questions according to the taxonomy of wh- words, there is a suggestion of a Zipfian distribution. Some possible explanations for the ubiquity of Zipfian distributions in the results of this study are discussed in chapter 5.

Table 4-11: Questions Classified According to the Taxonomy of Forms of Expected Answers

Question Classes	Percentage of Questions Classified
Ready reference	29.8
Readers advisory	22.2
Research	17.4
Analysis	11.7
Critique	4.0
Exact reproduction	4.0
Directional	3.5
Description	3.0
Citation list	0.3
Holdings	0.3
Bibliographic instruction	0

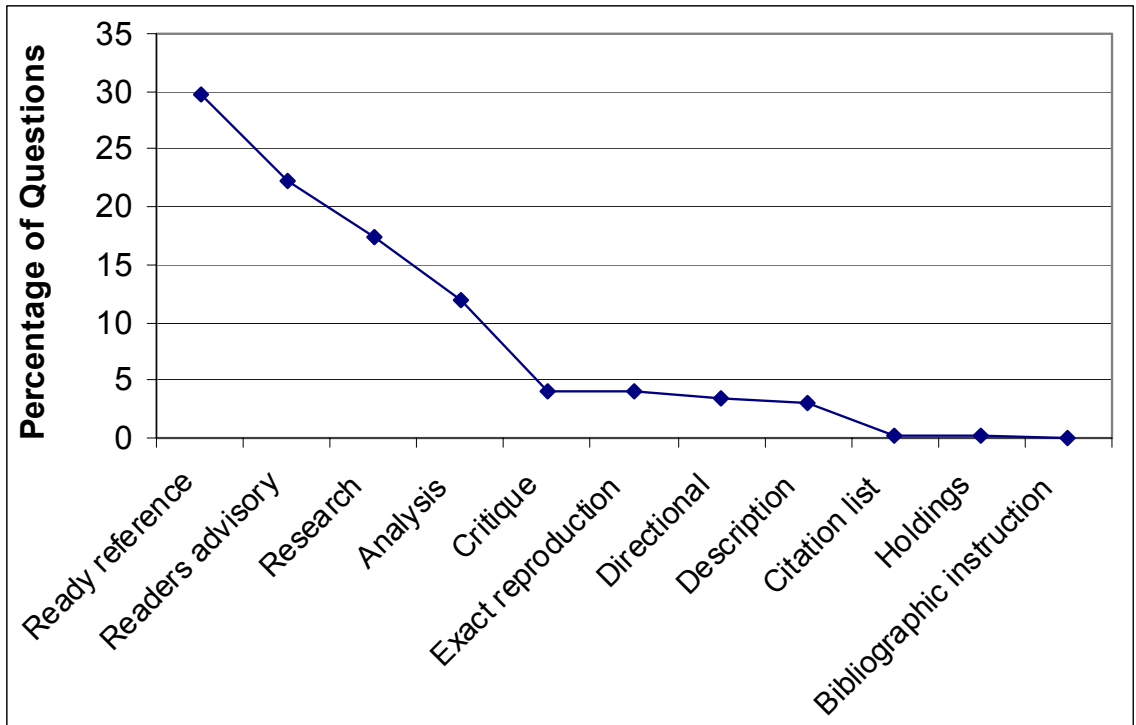


Figure 4-11: Distribution of Questions Classified According to the Taxonomy of Forms of Expected Answers

The scope of each of the classes in this taxonomy was somewhat ambiguous because, while the classes in this taxonomy have been discussed in the library literature for decades, they have never been consistently defined. Classifying questions according to this taxonomy, therefore, required more subjective judgment than classification using the other two taxonomies.

It seems therefore counterintuitive that this taxonomy should contain the smallest percentage of unclassifiable questions: 3.5%. This is not as counterintuitive as it seems, however, for the following reason: because classification according to this taxonomy is highly heuristic, questions cannot be classified unless the context is understood. In other words, the type of reference service makes a difference to the use of this taxonomy. For example, a question that might be considered ready reference if asked in a medical library might require a great deal of research if asked in a public library. (In this phase of the

study, the context was a digital reference service that provides general reference, and that does not possess its own collection, using the Internet as its primary reference collection.) Both the questions that users are likely to ask of a service, and the classification of those questions, are affected by the context of the service. For example, if a user understands that the service does not possess a collection, then it makes little sense to ask a Holdings question. On the other hand, it makes sense for a user who is unaccustomed to using the Internet as a reference source to ask a service a Readers' advisory question, as these questions ask for assistance in the selection of information sources. And indeed, 22.2% of questions were classified as Readers' advisory questions according to this taxonomy, such as the following: "I'm having great difficulty locating a biography on Judith Krantz... Would you be able to get me started in the right direction?" and "I need help on a Biology essay. It's all about heart disease and needs to include: Causes, Cures, and Preventions. I would be really grateful if you could advise me on some websites to look for reliable information."

4.3.1.3.1. Intersection of the Taxonomies of Forms of Expected Answers and Wh- Words

There is a fine distinction to be made between the Readers advisory and Directional classes when classifying questions submitted to digital reference services (as opposed to desk reference services), as Directional questions do not ask for assistance in the selection of information sources, but rather for the location of a specific source. Not surprisingly, there is a great deal of overlap between the Readers advisory and Directional classes in this taxonomy and the Where class in the taxonomy of wh- words: 50% of Readers advisory and 92.9% of Directional questions were classified as Where questions according to the taxonomy of wh- words.

Although containing only 11.9% of the questions classified according to this taxonomy, it is worth breaking down the distribution of questions within the Analysis class. Analysis questions intersect with the taxonomy of wh- words as follows: 36.2% were classified as How questions, 31.9% as What questions, and 21.3% as Why questions. This distribution

makes intuitive sense, in that questions that require analysis to answer should also employ a wh- word that is used to phrase open-ended questions. This distribution also holds true at the intersection with the taxonomy of functions of expected answers: 70% of Analysis questions were classified as classes that call for long answers.

4.3.1.3.2. Intersection of the Taxonomies of Forms of Expected Answers and Functions of Expected Answers

On the other side of the coin, it makes intuitive sense that Ready reference questions – a class for which the scope note states that they are “questions asking for simple, factual answers” – would call for short answers. So it proved to be: 88% of Ready reference questions were classified as classes that call for short answers according to the taxonomy of functions of expected answers.

A surprising finding is the distribution of questions in the Research class. It might be expected that Research questions would require long answers, as the scope note for this class explicitly states that such questions ask for “involved answers.” This turned out frequently not to be the case, as 61% of the questions in the Research class were classified as classes that call for short answers according to the taxonomy of functions of expected answers. Two examples of such questions were as follows, in bold:

“Why can’t we print enough money for everyone to be rich? **Who decides how much money we print** and how often do they decide that?”

“What is a Kalidascope [*sic*] and **who invented it?**”

The process of conducting the research necessary to answer these questions was likely quite involved, probably requiring some effort and perhaps the use of multiple information sources to answer. (The possibility of classifying questions according to their difficulty is discussed in chapter 5.) The answers, however, will both be short: the answer to the first question is “the U.S. Department of the Treasury” or “Secretary of the

Treasury Paul H. O'Neill,” and the answer to the second question is “Sir David Brewster.”

4.3.1.4. The Intersection of All Three Taxonomies

Intuitively it might be expected that the intersection containing the largest number of questions would be the intersection of the three classes containing the largest number of questions in each of the three taxonomies: What, Concept completion, and Ready reference. This is indeed the case, with 7.3% of the total number of questions in the data set. Examples of questions of this type are: “What is a synonym for ‘mountain making’?” and “What did ancient Egyptian [sic] people eat?” A close second, containing 6.6% of the total number of questions, is the intersection of How, Quantification, and Ready reference. Examples of questions of this type are: “How many counties are in the United States?” and “How many different cases of reported cowtipping have been reported in the last three months?”

On the other side of the coin, it is enlightening to also look at those intersections that contain few or no questions. In this phase of the study, prior to these taxonomies being modified, the taxonomy of wh- words contained 7 classes, the taxonomy of functions of expected answers contained 16 classes, and the taxonomy of forms of expected answers contained 10 classes. There are therefore $8 \times 11 \times 17 = 1,496$ “cells” in the “space” defined by these three taxonomies. There are, however, only 133 cells that contained questions (8.9% of the total number of cells), and only 60 cells containing more than one question (4%).

It was difficult, after this phase of the study, to determine the cause of this “clumpiness” of questions in the taxonomy space. One possibility is that it is an artifact of constraints on the range of ways in which it is possible to phrase a question in English – for example, the intersection of the When and Definition classes may simply be one in which questions will never exist, because a question phrased in this way make no sense in English (When is the meaning of X?). On the other hand, this clumpiness may be an artifact of the

reference service from which the questions in the data set were drawn. For example, the Virtual Reference Desk may have received so few Citation list questions because it is a service that answers questions directly; a service like AskERIC, on the other hand, which does not answer questions directly but instead provides lists of citations to documents, may receive more Citation list questions.

4.3.2. Intercoder Reliability

In order to test the reliability with which questions could be classified according to these taxonomies, as well as to test the clarity of the taxonomies and scope notes, an intercoder reliability test was performed. The statistic used to calculate the intercoder reliability was Cohen's κ (Cohen, 1960). Values for κ range between 1 and -1 , where 1 = perfect agreement beyond chance, 0 = no agreement beyond chance, and negative values = agreement worse than chance. Carletta (1996) states that $\kappa > 0.8$ is a good reliability measure, and $0.67 < \kappa < 0.8$ allows "tentative conclusions to be drawn" (p. 252). Uebersax (1987), however, claims that any attempt to quantify levels of agreement is a misuse of κ , and that κ should instead be considered to be a binary statistic: agreement either is or is not greater than what would be expected by chance. In either case, as can be seen in Table 4-12, the values of κ are better than what would be expected by chance. If one subscribes to Carletta's ranking of κ values, then one could make the conclusion that the values of κ for the taxonomy of wh- words indicate good reliability between the researcher and the volunteer coders, and fair reliability among the volunteer coders. The values of κ for the other two taxonomies are within the "tentative conclusions" range – the taxonomy of functions of expected answers on the high end of that range, and the taxonomy of forms of expected answers on the low end of that range. In addition, the values of κ between the researcher and volunteer coders are higher for all three taxonomies than the values of κ among volunteers.

Table 4-12: Phase 2 Intercoder Reliability Kappa Values

	Between researcher and volunteers	Among volunteers
Wh- words	$\kappa = 0.84$	$\kappa = 0.79$
Functions of expected answers	$\kappa = 0.75$	$\kappa = 0.72$
Forms of expected answers	$\kappa = 0.69$	$\kappa = 0.62$

These results are not particularly surprising. The taxonomy of wh- words is quite simple and intuitive for English speakers; it is therefore easy to understand how intercoder reliability for this taxonomy would be high. The taxonomy of functions of expected answers is less intuitive, but makes up for this to a degree by being quite precise – the scope notes for the classes in this taxonomy were derived from Lehnert’s (1978) and Graesser and colleagues’ (Graesser, Lang, and Horgan, 1988; Graesser, Person, and Huber, 1992; Graesser, McMahan, and Johnson, 1994) descriptions. These descriptions have been refined over time and through use, and therefore may be expected to be clear. The taxonomy of forms of expected answers is the least well-developed taxonomy of the three used in this study – while the classes in this taxonomy have been discussed in the library literature for decades, they have never been consistently defined.

It is less clear why the values for kappa among the volunteer coders are consistently lower than the values for kappa between the researcher and the coders. This may be a result of bias due to the effect of the researcher – the researcher created the scope notes for the taxonomies and gave instructions to the coders, but the coders did not communicate amongst themselves; thus the coders’ coding may reflect the researcher’s interpretation of the scope of the individual classes. In other words, the volunteer coders were trained in the use of these taxonomies, though inadvertently.

Two wrinkles exist in the analysis of this data, borne out of the content analysis issues discussed in section 3.2.5.2. First, a single email message may contain more than one question. This was a minimal problem, as there was an average of 89% agreement between the researcher and all volunteers as to how many questions there were in a single email. That is, for 89% of the email messages given to the volunteer coders, the researcher and the coders agreed (between the researcher and the coders, as well as among the coders) on how many questions were contained in any given email. Emails in the data set contained an average of 1.54 questions. This finding of the number of questions per email is consistent with Hert's (2000) average of 1.45 questions per email message received by the Bureau of Labor Statistics' web site in 1997.

In order to accurately calculate κ , it was necessary to ensure that the classifications for the same question were being compared. The instructions given to the volunteer coders instructed them to indicate the different questions that they identified in an email message (for example, by circling and numbering them), and classify each question independently (see Appendix C for the full text of these instructions). Thus it was possible to accurately match the questions classified by the researcher and the volunteer coders, in order to accurately calculate κ .

The second wrinkle in the analysis of this data is that the classes in the question taxonomies were not treated as being mutually exclusive – in other words, a single question may be of more than one question type. The instructions given to the volunteer coders instructed them to classify a question in more than one class if they believed that to be appropriate. In such cases, coders were instructed to list all of the classes to which they classified a question. This too was a minimal problem, as no volunteer coder had more than three questions that he or she classified in more than one class – and given that a single email message may contain more than one question, that is less than 0.15% of the questions that were multiply classified. Additionally, in all (100%) of these cases, one of the classes assigned to the questions by the volunteer coders were the same as classes assigned by the researcher.

The values of κ presented in Table 4-12 are good, but not great. These values indicate that the process of classification using these taxonomies is replicable. However, these values also indicate that the process of classification using these taxonomies could be improved. This improvement was accomplished by more rigidly defining the scopes of the classes, clarifying boundary objects, and the revising the scope notes to more clearly reflect these changes. This improvement is evident from the higher values of κ in the intercoder reliability testing performed after the classification of questions in Phase 3 of the study.

4.3.3. Evaluation of Question Taxonomies

Evaluation of the taxonomies of questions was performed after the classification of the data set of questions using these taxonomies. Each taxonomy was evaluated according to the set of thirteen criteria discussed in section 2.7.6. These criteria are presented in Table 4-13, below.

These criteria are all heuristic, serving as qualitative measures. These criteria were utilized in this phase of the study during the process of classifying the questions in the data set – that is, the three taxonomies used in this study were evaluated according to these thirteen criteria for classifying actual questions, during the process of classifying those questions.

4.3.3.1. Evaluation of the Taxonomy of Wh- Words

The scope of this taxonomy is the domain of all questions that it is possible to construct in English, according to the framework of the “five Ws.” The very simplicity of these rules causes the Partitioning criterion to be met for this taxonomy, as the difference between wh- words in English is a significant one. The five Ws, however, allow only for the construction of direct questions; they do not allow for the construction of indirect questions, such as requests stated as assertions (as the Assertion class in the taxonomy of functions of expected answers).

The fact that the five Ws can only accommodate direct questions is in part why there were such a large number of unclassifiable questions in this data analysis. This taxonomy is exhaustive in its coverage of the domain of direct questions, but lacks the ability to classify indirect questions of any sort. The simplicity of this taxonomy is both a strength and a weakness: questioners may find it natural to phrase questions according to the five Ws framework, but this taxonomy is unable to classify questions not phrased according to this framework. Of the three taxonomies employed in this study, this one may be the least useful, for two reasons. First, because it is unable to classify questions not phrased using a wh- word, it will inevitably require that a sizeable percentage of questions are unclassifiable. Second, it is the most obvious taxonomy: it is a simple matter to identify the wh- word in a question, or to determine that a question does not contain a wh- word. Given the simplicity of the analysis required to classify questions according to this taxonomy, doing so provides the least insight into the question.

The vocabulary of this taxonomy is highly coherent, with a consistent granularity, as the class names are simply the wh- words commonly used in English to construct questions. Due to the familiarity of the classes, the taxonomy is highly predictable, as well as easy to use. This familiarity also allows for usability, as probably all English speakers are familiar with the five Ws from an early age. The down-side of this familiarity, however, and one that detracts from the usability of this taxonomy, is that it is easy to forget that questions can be formed in English that do not utilize one of the five Ws (or the H).

There are two situations under which this taxonomy is particularly prone to problems: incorrect grammar on the part of the questioner, and the use of indirect questions. It was sometimes the case that the wh- word used in a question is not the appropriate one for the question that the questioner means to ask. For example, consider the question “In what literary work does [the phrase ‘It was the best of times, it was the worst of times,] appear?” This question uses the word “what,” which, according to the scope notes for this taxonomy (not to mention the rules of proper English usage) may only be used for infinite or undefined sets. For this question, the word “which” would have been more

appropriate, given that this question asks for the identification of a particular entity out of a finite set of entities (the universe of literary works). The inability of this taxonomy to correctly classify questions phrased grammatically incorrectly causes trouble for its Prototype characteristics; that is, a question may be classified in one class when it should be classified in another.

As stated above, this taxonomy is capable of classifying only direct questions, and does not allow for the classification of indirect questions. Unfortunately for this taxonomy, many “questions” submitted to digital reference services are not phrased as questions, but rather are phrased as requests or even assertions. For example, the question “Will you please give me info on patton maybe some pictures some websites I can go to” is phrased as a request, while the question “Tell me information about socretes [*sic*] an astrologer” is phrased as a directive. The question “I need some info on a mathematician by the name of Ellen Amanda Hayes” is phrased as an assertion. On the other hand, the question “Everything about garlic especially its medicinal applications” is not even a complete sentence, though it gets its point across effectively enough.

In conclusion, the taxonomy of wh- words serves the function of classifying questions phrased using the “five Ws” (and an H) very well. It is simple, intuitive, and easy to use. However, each class covers a very large area. In modifying this taxonomy, therefore, classes were divided into sub-classes where possible. Additionally, this taxonomy can only classify direct questions. In order to classify indirect questions, another taxonomy must be used.

4.3.3.2. Evaluation of the Taxonomy of Functions of Expected Answers

The scope of this taxonomy is the domain of all functions that it is possible for an answer to a question to have in fulfilling the questioner’s information need. This is not actually a taxonomy of information needs, but of fulfillments of information needs. While this taxonomy is exhaustive in its coverage of that domain, it has one major drawback: it is difficult to judge whether the dimensions along which classes are differentiated in this

taxonomy are significant or not. Dervin (1983) discusses what she calls helps and hurts, which she defines as “the uses made of information” (p. 17); these helps and hurts are used in her 1983 study and other studies (see for example Dervin et al., 1976; Dervin, Nilan, and Jacobson, 1982) to classify individuals’ fulfillments of information needs. The taxonomy of functions of expected answers, on the other hand, has not received similarly extensive use and validation for this purpose. Nevertheless, this taxonomy works as a scheme for classifying questions, and – significantly – across environments: in a laboratory setting (Lehnert, 1978), in teaching settings (Graesser, McMahan, and Johnson, 1994), at a library reference desk (White, 1998), and in a digital reference service (this study).

The class containing the largest percentage of the questions in the data set is Concept completion, containing 31.6% of the questions classified. While this class encompasses questions of the types Who, What, Which, When, and Where from the taxonomy of wh- words, the taxonomy would be more expressive if the Concept completion class had subclasses for each of these types. This taxonomy would be hospitable to these new classes by limiting the scope of existing classes to accommodate the scopes of the new classes.

The granularity of this taxonomy is inconsistent as well. While the Concept completion class contains questions that are phrased using several different wh- words, there are a number of classes that contain questions that are phrased using the word What in different ways: Feature specification, Quantification, Interpretation, and others. It is therefore difficult to determine what the defining characteristics of the prototypical entity in these classes are. For all that, however, this taxonomy is remarkably predictable and coherent in its vocabulary, as the classes are quite well defined in their scope. This makes this taxonomy quite easy to use.

4.3.3.3. Evaluation of the Taxonomy of Forms of Expected Answers

The scope of this taxonomy is the domain of all forms that it is possible for an answer to a question to take, in the context of a reference transaction. This is perhaps the most

difficult taxonomy according to which to classify questions: it requires a great deal of world knowledge of libraries and the provision of reference service. The classes in this taxonomy have never been well defined, and classifying questions according to this taxonomy requires more subjective judgment than classifying questions using the other two taxonomies. This taxonomy is therefore difficult to use and not particularly usable, as the scope of classes are somewhat ill-defined.

On the other hand, all three of the criteria concerned with the domain of the scheme are well-met by this taxonomy – as might be expected, given that this taxonomy is the result of decades of research and development in the measurement and evaluation of reference services. For this same reason, this taxonomy is highly consistent, at least for those familiar with the literature on library reference. Similarly, both of the criteria concerned with the domain of the classes themselves are met: the scopes of the classes are sufficiently broad that they can be subdivided into subclasses, and the taxonomy can easily accommodate new classes. The consistency and coherence of this taxonomy are less consistent, however, due to the fact that the scopes of the classes are not rigidly defined, and require some subjective interpretation. Likewise for the three criteria concerned with the relationships between classes and entities in classes.

4.3.3.4. Summary of the Evaluation

Table 4-13 summarizes the evaluation of the three taxonomies discussed in the previous sections.

Table 4-13: Thirteen Criteria for the Evaluation of Classification Schemes

Criteria Concerned with	Criteria	Wh- word	Functions of Expected Answers	Forms of Expected Answers
The domain of the entire classification scheme	Scope	+	+	+
	Exhaustivity	+	+	+
	Expressiveness	+	+	+
The domain of classes	Granularity	+	+	+
	Hospitality	–	+	+
Relationships between classes and entities in classes	Structure	+	+	+
	Partitioning	+	+	+
	Prototype characteristics	+	+	+
The terminology used in the classification scheme	Vocabulary	+	+	+
	Coherence	+	+	+
	Consistency	+	+	+
The utility of the classification scheme	Usability	+	+	–
	Browsability	+	+	–

This summary uses a binary division to indicate how each taxonomy “measures up” for each evaluation criterion. This pass-fail evaluation (pass = +, fail = –) is meant only to indicate the success of a taxonomy according to each evaluation criterion; it is not meant to indicate any sort of score. For example, for a taxonomy to receive a “pass” on the Scope criterion means that it covers the domain thoroughly, and entities are evenly distributed across classes. For a taxonomy to receive a “fail” on the Scope criterion means that it does not adequately cover the domain, and some classes contain far more

entities than other classes. These measures are highly subjective, based on the researcher's and volunteer coders' experiences in using these taxonomies to classify the questions in the data set. An interesting avenue for future research would be to formalize this evaluation of evaluation criteria, to provide a "meta-evaluation" framework for the evaluation of classification schemes.

4.3.4. Modification of Question Taxonomies

Based on the criteria used to evaluate the three taxonomies, discussed above, modifications were determined to be appropriate for the three taxonomies. These modifications were implemented prior to the classification task in Phase 3 of the study. These modifications are discussed here.

4.3.4.1. Modifications to the Taxonomy of Wh- Words

Robinson and Rackstraw (1972a) divide How questions into two sub-classes: How-method and How-process questions (p. 37). The distinction here is a fine one, and, as Robinson and Rackstraw state, may be dependent on the context in which the question is asked. In general, however, How-method questions ask about enablement (how is it possible to perform an action), while How-process questions ask about procedure (what are the steps involved in performing an action). Based on Robinson and Rackstraw's differentiation of these types of questions, two modifications were made to this taxonomy: 1) the class What-description was added as a sub-class of What questions, to encompass questions about procedure, and any other type of question that requires that the answer take the form of a description or explanation of one or more entities or actions, and 2) the scope notes for the How class was modified to encompass questions about enablement.

Hovy, Hermjakob, and Lin (2001) suggest the class of How-quantity questions: for example, questions phrased using "how many," "how much," and "how long." It is only because of the grammatical idiosyncrasy of the English language that this class of questions does not already exist – in Spanish, for example, there are two different wh-

word-equivalents that both translate into English as How: “cómo” translates to “how,” as in “how is it,” while “cuánto” translates to “how much.” At Hovy, Hermjakob, and Lin’s suggestion, a class with this scope was also added to this taxonomy: the What-quantity class. This class was added as a sub-class of What questions, rather than of How questions, because questions of this type require the specification of a state or value out of the set of all possible values (the scope of What questions), rather than the description or an evaluation of a process or procedure (the scope of How questions). This class will encompass questions that require a quantitative description of the state or value of attributes of objects or events.

These two sub-classes (What-description and What-quantity) were added to the taxonomy of wh- words. What-selection was also added as a sub-class of the What class, to encompass the more traditional what-type questions: questions that require the selection of one or more members of a (specified or implied) infinite set.

One additional issue was addressed in the modifications to this taxonomy that could not be addressed through modifications to any of the classes, but rather was addressed through a scope note for the entire taxonomy: the issue of questions phrased idiomatically. For example, “how come” and “what for” are idiomatic ways of saying “why.” The scope notes for this taxonomy, after modifications were made, as it was used in Phase 3 of the study, instructed the coders that just because a question is phrased using a particular wh- word, does not mean that it necessarily belongs in the class named for that wh- word.

4.3.4.2. Modifications to the Taxonomy of Functions of Expected Answers

As has already been discussed, this taxonomy is the most fully developed of the three utilized in this study. It therefore needed little modification, as the scopes of the classes were already clear, and the scope of the taxonomy as a whole was nearly comprehensive.

Two classes were added to this taxonomy, that were warranted by the existence of questions of a particular type in the data set. These two classes were named Explanation and Coverage. The Explanation class is similar in scope to the Why class in the taxonomy of wh- questions, in that it encompasses questions requiring a justification of an object or event. For example, some questions that were unclassifiable according to this taxonomy that could be reclassified in the Explanation class are: “Why does the thickness of a string affect the sound it makes?” and “Why are costlines [*sic*] seldom straight?” The Coverage class is similar in scope to the Readers advisory class in the taxonomy of forms of expected answers, in that it encompasses questions that demand materials on a subject specified. For example, some questions that were unclassifiable according to this taxonomy that were reclassified in the Coverage class are: “please send me information on WW2 for wartime in Britain” and “What is the state of converting conventional vehicles to renewable energy technology (Like Fuelcell) and hybrid electric [*sic*].”

The scopes of the Instrumental/procedural and the Enablement classes were modified to clarify the split between these two classes that are concerned with an agent’s performing an action or accomplish a goal. First, the Instrumental/procedural class was renamed Procedural to indicate that questions of that class require the specification of a procedure. The Enablement class, then, requires the specification of an object or resource. The scope notes for both of these classes were modified to indicate this split.

Other than these two added classes, the only modifications to this taxonomy were simple name changes and the creation of the beginning of a hierarchical structure. To insure that the names of classes are descriptive and as clear as possible, the Concept completion class was renamed Fill-in-the-blank, and the Request/Directive class was renamed Request. Additionally, a super-class named Causality was added to this taxonomy, encompassing questions requiring a causal description or explanation of states or events. Under the Causality class fell the sub-classes Causal antecedent, Causal consequent, and Expectation. Another super-class named Questions phrased as statements was also added. Under this class fell the sub-classes Assertion and Request.

4.3.4.3. Modifications to the Taxonomy of Forms of Expected Answers

The taxonomy of forms of expected answers is the least well-defined of the three employed in this study. There was no consistency in the unclassifiable questions, to warrant the creation of new classes based on the questions in this class. There was, however, a distinction among questions in the Ready reference class, which warranted the creation of a new class: Known-item search. The Known-item search class contained questions requiring a specific datum that the questioner indicates (explicitly or implicitly) that he or she knows exists. For example, a question that was classified in the Ready reference class that could be reclassified in the Known-item search class is: "I have been searching the web looking for someone to tell me what national holiday is referred to as Founder's Day? Is it President's Day? I remember seeing ads in the newspapers touting Founder's Day sales but I cannot remember what holiday it is."

Other modifications to this taxonomy involved clarifying the scopes of classes. The most dramatic of these modifications was the elimination of the Research class. The scope note for the Research class stated that it encompassed questions asking for involved answers, requiring some effort and wide use of information sources. This scope was not appropriate for this taxonomy, as it defined the class in terms of the effort required to answer a question, rather than the form that the answer should take. The scopes of the Analysis and Critique classes were therefore expanded to encompass questions that might formerly have been classified as Research questions.

For the sake of clarity, three classes were renamed: Ready reference to Factual, Exact reproduction to Reproduction, and Description to Summary. The name Ready reference is highly contextual in reference work: a question that might be considered ready reference if asked in a medical library might require a great deal of research if asked in a public library. To avoid the issue of context for this class, it was renamed Factual and the scope note modified to encompass any question requiring facts. The Exact reproduction

class was renamed Reproduction to allow for requests for reproductions of either an entire information source or a subset of it. The Description class was renamed Summary and the scope note modified to encompass questions requiring any form of summary of an information source, not only a description.

4.4. Phase 3: Classification Task

The classification phase of this study addressed Research Question 2:

RQ2. How does question type correlate with the action taken on a question in the triage process?

This section discusses the findings from the classification phase of this study, and answers Research Question 2. Values for the intercoder reliability statistic Cohen's κ (Cohen, 1960), were computed between different sets of coders. Values of κ covered the range 0.53 to 1, indicating weak to perfect agreement. Values for the correlation statistic Cramér's V (Cramér, 1966) for the correlation between question taxonomies and question attributes ranged from 0.06 to 0.27, indicating weak correlation. Values for V for the correlation between question taxonomies and triage actions ranged from 0.15 to 0.71, indicating some weak and some moderately strong correlations. Some very strong correlations were found between specific intersections of question types and triage action.

4.4.1. Classification of Questions

This section describes the findings from the classification of the set of thirty questions sampled from the pool of 185 questions collected during the think-aloud phase of this study. As in the classification task in Phase 2 of this study, this classification was performed according to the three taxonomies of questions identified in the literature. Unlike in Phase 2, however, this classification was performed according to the three taxonomies of questions in their form after the modification performed at the conclusion of Phase 2. This phase of the study was conducted in order to classify a subset of the

questions collected during the think-aloud phase of this study, in order that the correlation between question type and triage action could be computed on these collected questions.

Nine coders were solicited from the pool of twenty-eight digital reference triagers who were respondents in the think-aloud phase of this study. These nine coders were provided with a set of thirty questions sampled from the pool of 185 questions collected during the think-aloud phase of this study, the three taxonomies of questions and scope notes for the taxonomies, and a set of instructions. (See Appendix D for the full text of these instructions.) The coders were instructed to classify the questions provided to them according to the three question taxonomies, using the scope notes for the three taxonomies as a guide.

4.4.1.1. Taxonomy of Wh- Words

Table 4-14 lists and Figure 4-12 shows the distribution of the set of thirty questions sampled from the pool of 185 questions collected during the think-aloud phase of this study classified according to the taxonomy of wh- words.

Table 4-14: Questions Classified According to the Taxonomy of Wh- Words

Question Classes	Percentage of Questions Classified
What-description	20
What-selection	20
Where	16.7
How	13.3
When	6.7
What-quantity	3.3
Which	3.3
Why	3.3
Who	0

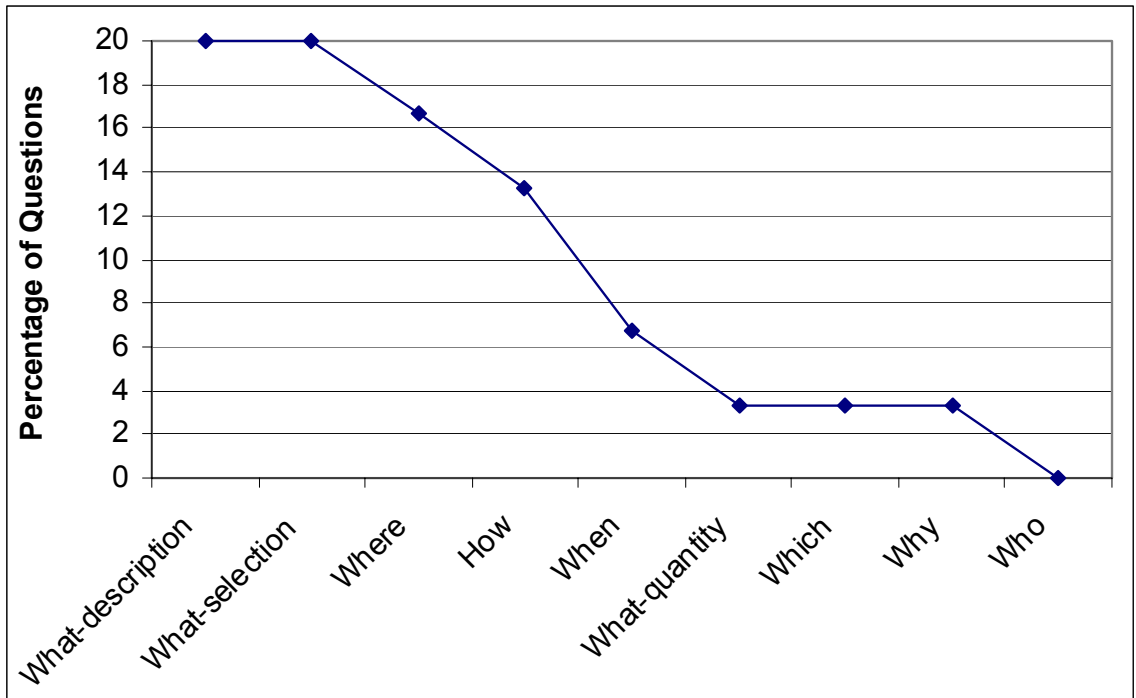


Figure 4-12: Distribution of Questions Classified According to the Taxonomy of Wh- Words

As in the distribution of questions classified according to the taxonomy of wh- words in Phase 2 (see section 4.3.1.1), there was a large percentage of unclassifiable questions (13.3%). The reasons that questions were unclassifiable are discussed below.

An interesting fact to note about the distribution shown in Figure 4-12 is that it is one of the few distributions in this study that does not follow a Zipfian distribution; rather, this distribution follows a fairly smooth decline. The reasons that so many distributions in this study follow a Zipfian distribution are discussed in chapter 5. It is difficult to determine why this particular distribution did not follow a Zipfian distribution. It is possible that it is simply an artifact of there being so few classes in this taxonomy (10) and so few questions classified (30); if there were more data points it is possible that this distribution would resemble a Zipfian distribution.

4.4.1.2. Taxonomy of Functions of Expected Answers

Table 4-15 lists and Figure 4-13 shows the distribution of the set of thirty questions classified according to the taxonomy of the functions of expected answers.

Table 4-15: Questions Classified According to the Taxonomy of Functions of Expected Answers

Question Classes	Percentage of Questions Classified
Coverage	30
Request	16.7
Quantification	10
Verification	10
Explanation	6.7
Procedural	6.7
Causal consequence	3.3
Comparison	3.3
Definition	3.3
Disjunctive	3.3
Feature specification	3.3
Judgmental	3.3
Assertion	0
Causal antecedent	0
Enablement	0
Example	0
Expectation	0
Fill-in-the-blank	0
Goal orientation	0
Inference	0

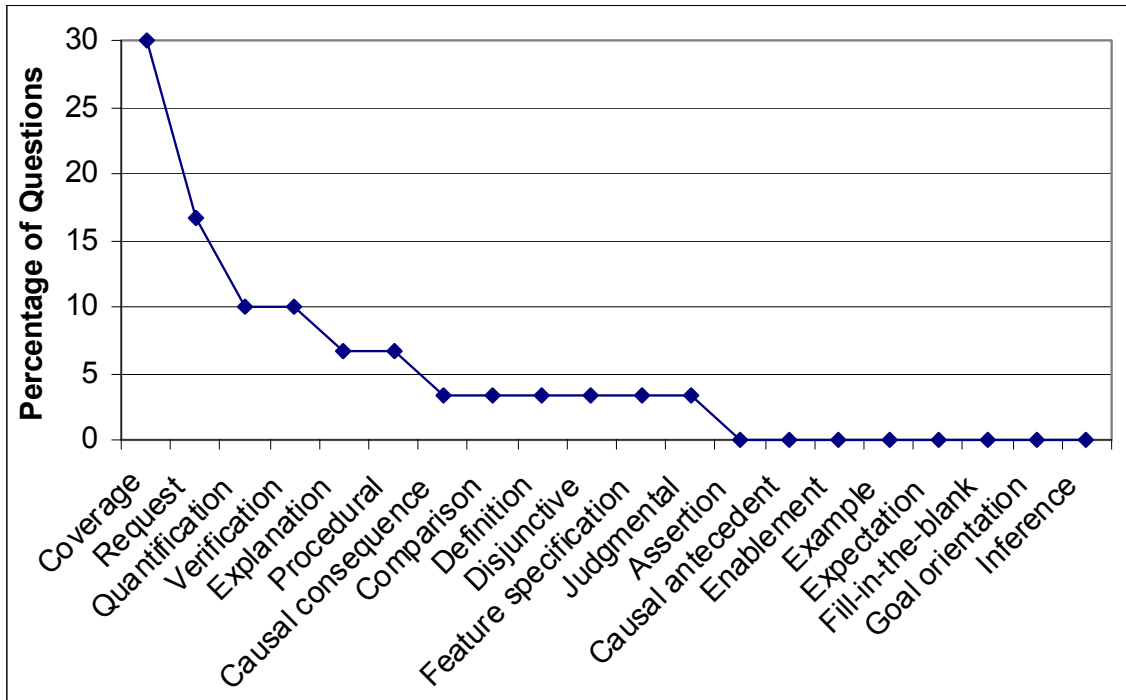


Figure 4-13: Distribution of Questions Classified According to the Taxonomy of Functions of Expected Answers

The class containing the greatest percentage of questions was Coverage, at 30% of the questions classified. The Coverage class was one of the two classes that was added to this taxonomy at the conclusion of Phase 2, and includes questions that demand materials on a subject specified. This class encompasses a common type of question in reference service, the question of the form, “please send me information on X,” or “I’m looking for information on X.” Thus it is not surprising that the most questions were classified according to this class.

As in the distribution of questions classified according to the taxonomy of the functions of expected answers in Phase 2, the Request, Quantification, and Verification classes contain the second, third, and fourth largest percentages of questions. The fact that the distribution of questions according to this taxonomy remains stable regardless of the set of questions classified is an indication of the validity of this taxonomy.

4.4.1.3. Taxonomy of Forms of Expected Answers

Table 4-16 lists and Figure 4-14 shows the distribution of the set of thirty questions classified according to the taxonomy of the forms of expected answers.

Table 4-16: Questions Classified According to the Taxonomy of Forms of Expected Answers

Question Classes	Percentage of Questions Classified
Factual	50
Citation list	23.3
Directional	6.7
Readers advisory	6.7
Analysis	3.3
Bibliographic instruction	3.3
Holdings	3.3
Known-item search	3.3
Critique	0
Reproduction	0
Summary	0

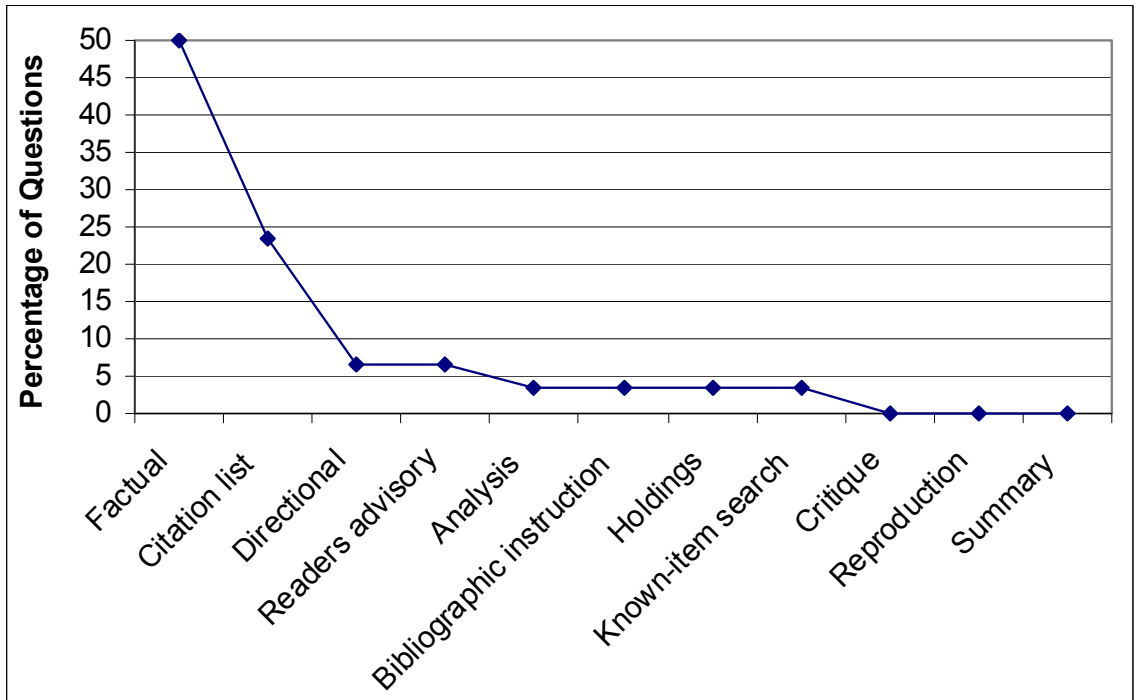


Figure 4-14: Distribution of Questions Classified According to the Taxonomy of Forms of Expected Answers

As in the distribution of questions classified according to the taxonomy of the forms of expected answers in Phase 2, there were few (in fact no) unclassifiable questions. The reasons that questions were unclassifiable are discussed below. But it is an indication that the scope notes for this taxonomy were clarified to the point where all questions classified could be accommodated by this taxonomy.

The distributions shown in Figures 4-13 and 4-14 follow a Zipfian distribution; the reasons that so many distributions in this study follow a Zipfian distribution are discussed in chapter 5.

4.4.2. Unclassifiable Questions

In the instructions provided to the coders for performing the classification task, there was a note explaining that they could classify a question as “Other” in any of the three

taxonomies, indicating that a question could not be classified according to any of the other classes in that taxonomy. Thus, it was possible for a coder to decide that a question did not fit neatly into any of the classes in a taxonomy, and that the question was unclassifiable according to that taxonomy.

Kwaśnik (1999) states that the sign of a ‘premature’ classification “is the need for a ‘miscellaneous’ or ‘other’ category into which the classifier places all those entities that do not fit into the logic of the classification system as specified” (pp. 28-29). Despite the undesirability of a Miscellaneous class in a taxonomy, it was necessary to include one in all three taxonomies utilized in this study. This was necessary because these taxonomies are still developing. These taxonomies existed and had been developed by others prior to this study, and were identified through an extensive review of the literature from several fields that deal with questions. Modifications were made to these taxonomies based on the evaluation of how the taxonomies performed at the conclusion of Phase 2, and prior to the beginning of Phase 3 of this study. Nevertheless, to allow for the possibility that even given the modifications to these taxonomies performed as a part of this study, these taxonomies are still not exhaustive in their coverage of their domain, the possibility of unclassifiable questions was included in the coders’ instructions.

As it turned out, very few questions were unclassifiable according to any of the three taxonomies: four according to the taxonomy of wh- words (13.3% of the 30 questions classified), and none according to the taxonomies of functions of expected answers and forms of expected answers. The four unclassifiable questions according to the taxonomy of wh- words were questions that at least one coder classified as Other, even if other coders classified the same question in a different class.

In the distributions of questions classified according to the three taxonomies in Phase 2, the number of unclassifiable questions decreased as the taxonomies go up in levels of linguistic analysis. In the distributions of questions classified according to the three taxonomies in this phase of the study, unclassifiable questions only appear in the taxonomy at the lowest level of linguistic analysis, the taxonomy of wh- words. This

finding could be interpreted in two ways: on the one hand, it may become easier to classify a question as an increasing amount of context surrounding the question is taken into account. On the other hand, it may be that the higher the level of linguistic analysis, the more exhaustively it is possible for a classification scheme at that level to cover the domain of speech acts, such as questions.

In either case, if only one taxonomy of questions could be utilized to classify questions from digital reference services, it is a toss-up between the taxonomy of functions of expected answers and the taxonomy of forms of expected answers, as coders were able to classify every question provided to them utilizing both of these taxonomies. Taking second place is the taxonomy of wh- words. The clear choices of taxonomies to utilize in designing systems to automate the triage process, the clearest and most unambiguous taxonomies: the taxonomies of functions of expected answers and forms of expected answers.

4.4.3. Questions on which Coders Disagreed

In order to discover the reasons why coders disagreed on the classification of questions, interviews were conducted with each of the three coders in a group separately, about the questions on which there was disagreement. These interviews were conducted in order to determine why coders classified those questions as they did. The reasons that were articulated by the coders were compared within the groups of three coders, and the underlying causes for these disagreements were sought. In performing these interviews, it was necessary to determine how many and which specific questions out of the sets of ten were disagreed on. This “raw disagreement,” as it were, is the inverse of the “raw agreement” referred to in section 3.3.4 as the variable A, the number of agreements between coders in their classification (Neuendorf, 2002, p. 149).

4.4.3.1. Taxonomy of Wh- Words

There were three questions on which coders 1, 2, and 3 disagreed on the classification according to the taxonomy of wh- words, and four on which coders 4, 5, and 6 disagreed. The taxonomy of wh- words encompasses only direct questions; it cannot represent indirect questions, such as requests stated as assertions (as the Assertion class in the taxonomy of functions of expected answers). Not surprisingly, some of the seven questions on which coders disagreed on the classification according to this taxonomy were phrased as indirect questions. Two of these indirect questions were: “name the Major River in Florida,” and “I am seeking journals/articles relating to student test preparedness behaviors.” Given that the taxonomy of wh- words is at the syntactic level of linguistic analysis, indirect questions like these simply cannot be encompassed by this taxonomy. This taxonomy analyzes questions according to the word that is used syntactically to form a question. Indirect questions do not use such words, because they are not phrased syntactically as questions. The “question” about Florida at least is phrased as a request or a demand, and as such possesses the same illocutionary force as a question, compelling a response from the hearer. The “question” about test preparedness behaviors, on the other hand, is simply a statement. This question has illocutionary force only in the context of a reference service, in which it is understood that when a patron says something like “I am seeking something,” that that should be interpreted to possess the same illocutionary force as a question.

Some of the seven questions on which coders disagreed, on the other hand, were direct questions, but simply not phrased using a wh- word. Two of these questions were: “Do you have any good information on creatine and weight lifting?,” and “I know that Spanish Missions in California tended to follow the coastline. Is there a similar pattern of coastal missions in Mexico?” Both of these questions, on their face, are questions requiring a yes/no answer – “Do you have...” and “Is there a...” – and as such might be interpreted to be Which questions. However, rephrasing these questions as Which questions would be the equivalent of answering “Yes” in response to the question “Do you know what time it is?” Indirect questions cannot be encompassed by this

taxonomy. Direct questions phrased without using a wh- word, on the other hand, *should not* be encompassed by this taxonomy: such questions *can* be encompassed by this taxonomy, but to do so loses the meaning of the question that can be understood by analyzing the question at a higher level of linguistic analysis.

4.4.3.2. Taxonomy of Functions of Expected Answers

There were four questions on which coders 1, 2, and 3 disagreed on the classification according to the taxonomy of functions of expected answers, four on which coders 4, 5, and 6 disagreed, and one on which coders 7, 8, and 9 disagreed. The questions on which coders disagreed according to this taxonomy were those that on their face were phrased as short-answer questions. Two such questions phrased as Verification questions, requiring a yes/no answer, were: “Do african american males/students learn better from african american teachers?” and “Is there a chart that can give you distance between one star and another.” Another question phrased as a Fill-in-the-blank question, requiring the determination of the subject of a sentence, was: “What enemies do hammerhead sharks have?” To treat these questions as being Verification or Fill-in-the-blank questions, however, is to ignore that they were asked in the context of a reference service, and thus are requests for more information – which can be understood by analyzing the question at a higher level of linguistic analysis.

4.4.3.3. Taxonomy of Forms of Expected Answers

There was one question on which coders 1, 2, and 3 disagreed on the classification according to the taxonomy of forms of expected answers, and one on which coders 4, 5, and 6 disagreed. These questions were: “I would need documentation on schools associated with Zoo’s, grades 7 – 12 and 2 – 6,” and “what is the U.S. government’s definition of urban education?” With a sample size of only two, it is difficult to come to any conclusions as to why coders disagreed on the classification of questions according to this taxonomy. Two coders classified the zoo question in the Citation list class, and one in the Factual class. The disagreement between the coders was

in the interpretation of the questioner's intention in asking the question: the two coders who classified this question as Citation list interpreted the question as being a request for *documents or citations to documents* about schools associated with zoos, while the coder who classified this question as Factual interpreted the question as a request for *data* about schools associated with zoos. Two coders classified the urban education question in the Factual class, and one in the Analysis class. Again, the disagreement between the coders was in the interpretation of the questioner's intention: the two coders who classified this question as Factual interpreted the question as being a request for a *definition*, perhaps reproduced from a government document, while the coder who classified this question as Analysis interpreted the question as a request for an *analysis* of urban education in contrast to education in other settings.

Disagreements between coders in classifying questions according to taxonomies at lower levels of linguistic analysis can be explained by insufficient context. The taxonomy of forms of expected answers, however, is at the pragmatic level of linguistic analysis – the highest level, taking into consideration “the purposeful use of language in situations, particularly those aspects of language which require world knowledge” (Liddy, 1998, p. 15) – so on the face of it, it seems unreasonable that disagreements between coders could be due to a lack of context. This is, however, precisely the case. The disagreements between coders in classifying questions according to the taxonomy of forms of expected answers were due to a lack of context surrounding the question: in other words, what was the questioner attempting to achieve in asking the question? What information need was he or she attempting to fulfill? What was the questioner's intended use of the information that would be provided? These are all elements of the context that were either missing or ambiguous in these questions, as they were provided to the coders. Many of these elements of context can be inferred from the question as it is received by a digital reference service – this is clear from the fact that there were so many fewer disagreements between coders for the taxonomy of forms of expected answers than for either of the other two taxonomies. Some elements of context, however, cannot, it appears, be inferred from the question as received. This is support – if any was required – for the utility of the reference interview in digital reference, or at least something

resembling it, in the form of reference interview-like questions on a question submission webform.

4.4.4. Intercoder Reliability

In order to test the reliability with which questions could be classified according to the modified taxonomies, an intercoder reliability test was performed. The statistic used to calculate the intercoder reliability was Cohen's κ (Cohen, 1960).

Possible values for the intercoder reliability statistic Cohen's κ range between 1 and -1 , where 1 = perfect agreement beyond chance, 0 = no agreement beyond chance, and negative values = agreement worse than chance (Cohen, 1960). Carletta (1996) states that $\kappa > 0.8$ is a good reliability measure, and $0.67 < \kappa < 0.8$ allows "tentative conclusions to be drawn" (p. 252). Uebersax (1987), however, claims that any attempt to quantify levels of agreement is a misuse of κ , and that κ should instead be considered to be a binary statistic: agreement either is or is not greater than what would be expected by chance. In either case, as can be seen in Table 4-17, the values of κ are better than what would be expected by chance. If one subscribes to Carletta's ranking of κ values, then one could make the conclusion that the values of κ for the taxonomy of forms of expected answers indicate good reliability for all three groups of coders. The values of κ for the other two taxonomies span the good and the "tentative conclusions" range.

One of the reasons that κ was selected as the statistic for measuring intercoder reliability for this study was that it may be used to determine agreement between more than two coders (Fleiss, 1971). Because the nine coders were divided into three groups of three, where each group classified the same questions according to the same taxonomies, intercoder reliability was computed between all three coders in each group. Table 4-17 presents the values of κ for these three groups.

Table 4-17: Phase 3 Intercoder Reliability Kappa Values

	Wh- words	Functions of expected answers	Forms of expected answers
Group 1	$\kappa = 0.75$	$\kappa = 0.61$	$\kappa = 0.89$
Group 2	$\kappa = 0.70$	$\kappa = 0.70$	$\kappa = 0.91$
Group 3	$\kappa = 1$	$\kappa = 0.92$	$\kappa = 1$

As shown in Table 4-12, above, the values of κ in the intercoder reliability test during Phase 2 ranged from 0.62 to 0.84. As shown in Table 4-17, the values of κ in this phase of the study cover a wider range – 0.61 to 1 – but the low end of that range is consistent with the Phase 2 results, and the high end is 1, indicating perfect agreement.

4.4.5. Correlation between Question Type and Triage Action

Once the set of thirty questions had been coded by the nine coders, the correlations between the three question taxonomies and the attributes of these questions that affected triagers' triage decisions, and between the three question taxonomies and triage actions were computed. The correlation statistic utilized for this was Cramér's V (1966). In addition, correlations between specific intersections of question types and triage actions were explored.

Because there was not perfect intercoder reliability in the classification of all questions, the correlation between question taxonomies and triage actions could not be computed with perfect validity for all questions. On the classification of some questions, all three coders agreed, and the correlation was computed between that question's class and the triage action taken upon it. On those questions for which two out of three coders agreed on the classification, the correlation was computed between the class that was agreed on by the two coders and the triage action taken upon it, but this correlation must be

interpreted as being only two-thirds reliable, given that only two out of three coders agreed on the classification. Fortunately, there were no questions for which all three coders disagreed; this avoided the problem of deciding for which class to calculate the correlation with triage action. This demonstrates that either the questions or the taxonomies and scope notes, or both, were clear.

Table 4-18 lists the values for Cramér’s V for the correlation between question taxonomies and question attributes that affect triage. These values are low, indicating weak correlations. The values of V are larger for questions on which two of the three coders agreed, perhaps indicating that there are stronger correlations between question type and question attribute when intercoder reliability is lower. On the other hand, this could simply be a reflection of the fact that there were fewer questions on which two of the three coders agreed, which could affect the values of a χ^2 -based statistic such as Cramér’s V. Correlations between question taxonomies and the other five categories of attributes (attributes of the answer, the patron, the triaging service, the receiving service, and the answerer) were equally weak.

Table 4-18: Values of Cramér’s V for the Correlation between Question Taxonomies and Question Attributes

	Question taxonomy	Cramér’s V
Questions on which all three coders agreed	Wh- words	V = 0.06
	Functions of expected answers	V = 0.07
	Forms of expected answers	V = 0.06
Questions on which two of three coders agreed	Wh- words	V = 0.11
	Functions of expected answers	V = 0.10
	Forms of expected answers	V = 0.27

Table 4-18 shows that there is a weak correlation between question taxonomies and the question attribute that affects the triage process. These findings indicate that the question attribute that affects the triage process is largely unaffected by the taxonomy that is used to classify the question. This is actually not very informative, as it is not the specific taxonomy used to classify questions that is of interest in answering Research question 2, but rather the specific question classes into which questions are classified.

Research question 2 asks, how does question type correlate with the action taken on a question in the triage process? Table 4-19 goes further towards answering Research question 2 by listing the values for Cramér’s V for the correlation between question taxonomies and triage action, according to the simple categorization of triage recipients discussed in section 4.2.2.5: whether the recipient is internal or external to the digital reference service or the organization with which the service is affiliated.

Table 4-19: Values of Cramér’s V for the Correlation between Question Taxonomies and Triage Action

	Question taxonomy	Cramér’s V
Questions on which all three coders agreed	Wh- words	V = 0.15
	Functions of expected answers	V = 0.23
	Forms of expected answers	V = 0.18
Questions on which two of three coders agreed	Wh- words	V = 0.38
	Functions of expected answers	V = 0.58
	Forms of expected answers	V = 0.71

Table 4-19 shows that the correlations between question taxonomies and triage action range from low to moderately high, indicating some weak and some moderately strong correlations. Again, the values of V are larger for questions on which only two of the

three coders agreed. These findings indicate that, depending on the taxonomy, the action taken on a question is weakly to moderately affected by the taxonomy that is used to classify the question. This goes further towards answering Research question 2, but the level of analysis is still the taxonomy, rather than the question class.

Table 4-19 lists the values for Cramér’s V for the correlation between question taxonomies and triage action, across all services. Remaining at the taxonomy level of analysis just a bit longer, Table 4-20 lists the values for Cramér’s V for the correlation between question taxonomies and triage action, for only the four digital reference services that received the highest volume of questions during the think-aloud studies (30 questions for three services, and 24 questions for one service). Because the think-aloud studies were conducted only up to a maximum of thirty questions, in order to avoid the problem of small sample sizes affecting the χ^2 -based statistic Cramér’s V, Table 4-20 collapses into one group the questions on which all three coders agreed and on which only two out of three coders agreed. What is most remarkable about Table 4-20 is the consistency of the values of V for each of the taxonomies, within each service. While the values of V indicate only moderately strong correlations between question taxonomies and triage action, the consistency in the values of V indicate that this correlation holds for all services, and regardless of the taxonomy utilized to classify questions.

Table 4-20: Values of Cramér’s V for the Correlation between Question Taxonomies and Triage Action, in the Four Highest-volume Services

Question taxonomy	Cramér’s V ₁	Cramér’s V ₂	Cramér’s V ₃	Cramér’s V ₄
Wh- words	0.58	0.45	0.44	0.35
Functions of expected answers	0.58	0.41	0.45	0.38
Forms of expected answers	0.58	0.41	0.44	0.37

Research question 2 asks, how does question type correlate with the action taken on a question in the triage process? The findings presented in Tables 4-18, 4-19, and 4-20 provide a partial answer to this question: moderately strongly, when correlations with triage action are computed utilizing the taxonomy as the level of analysis. Stronger correlations are evident, however, when the question class is utilized as the level of analysis: very strong correlations are found between specific intersections of question types and triage action. To interpret Table 4-21, below, imagine a “taxonomy space” defined by the three question taxonomies, as represented in Figure 4-15.

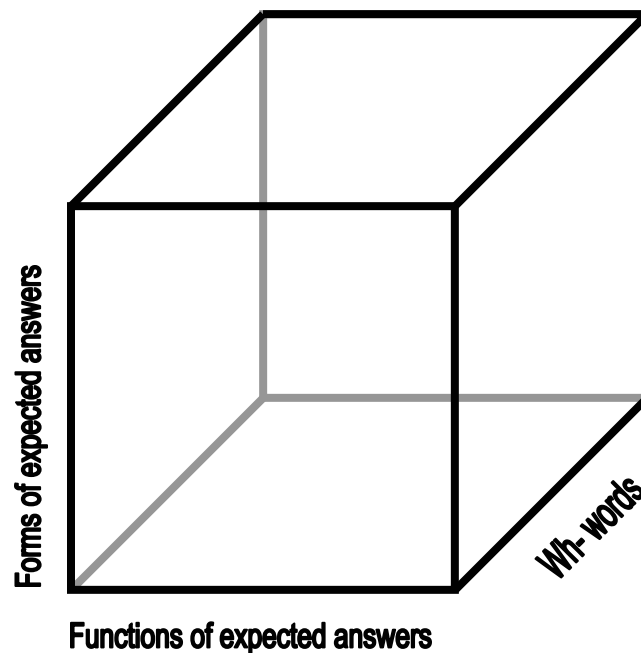


Figure 4-15: Taxonomy Space Defined by the Three Taxonomies

Each axis of the taxonomy space represented in Figure 4-15 is one of the three taxonomies utilized in this study, and the classes in each taxonomy fall along the appropriate axis. With ten classes in the taxonomy of wh- words, twenty-one in the taxonomy of functions of expected answers, and twelve in the taxonomy of forms of expected answers, there are 2,520 “cells” in this taxonomy space, and each of those cells may be thought of as a unique question type. It is not possible for questions to exist in

some of these cells, however, due to constraints on the range of ways in which it is possible to phrase a question in English: for example, there can be no questions at the intersection of the When and Definition classes, because a question phrased in this way make no sense in English (When is the meaning of X?). Thus, in practice, there are fewer than 2,520 cells in this taxonomy space in which questions can exist; a useful future direction for research would be to identify all of the cells in which questions can and cannot exist.

Table 4-21: Very Strong Correlations between Intersections of Question Types and Triage Action

Cell in the taxonomy space	Triage action	# of questions in cell	Example question
----- Coverage Factual	Internal	2	“Is there a chart that can give you distance between one star and another” “How were scalp diseases treated in the late 19th Century?”
----- Coverage Citation list	Internal	4	“Is there any system or body of experience in teaching bridge to school children?” “Do have research on including marshal arts, wresting and/or boxing modules in alternative school settings?”
----- Quantification Factual	Internal	2	“Do babies see in color right away, if not when do they?” “Would you be able to tell me how much money from the Gross National Product goes into the Military?”

What-description Request -----	Internal	3	“name the Major River in Florida.” “Can you tell me information about bronze items made in the Shang Dynasty?”
Where Request Factual	Internal	2	“Where was the first McDonald’s restaurant franchise location in Seattle?” “I am seeking journals/articles relating to student test preparedness behaviors. Do you have any suggestions as to where I might find this information?”
How Explanation Factual	Internal	2	“How does quinine work?” “I know that Spanish Missions in California tended to follow the coastline. Is there a similar pattern of coastal missions in Mexico?”
----- Coverage Directional	External	2	“Where can I find info on the Internet relating to the European Economy during 1400 to 1700 AD?” “What is the website for the SCANS Report?”
What-selection Coverage Citation list	External	2	“Need documentation on schools associated with zoos.” “Need lesson plans to assist with classes related to memorials for Sept. 11th.”

Table 4-21 presents the intersections of question types – cells within or slices through the taxonomy space – at which the correlation between question type and triage action is

100%. In other words, all of the questions that were classified at the specific intersections listed in Table 4-21 were triaged as indicated, according to the simple categorization of triage recipients discussed in section 4.2.2.5, whether the recipient is internal or external to the digital reference service. Different digital reference services triage questions to different sets of recipients, and this simple categorization of recipients was utilized in this phase of the study due to this wide range (the possibility of studying the range of triage recipients in one service only is discussed in chapter 5). Cramér's V was not computed for the correlation between these cells and triage action, because χ^2 -based statistics such as Cramér's V are used to compute the correlation between variables in 2 x 2 tables and larger, and one cell x triage action is a 1 x 2 table.

As discussed above, only thirty questions were classified during this phase of this study. The cells listed in Table 4-21 thus do not contain very large numbers of questions: most contain two, three, or four questions. It is likely that as the number of questions in each cell increases, the correlations between question type and triage action would remain very strong, though perhaps not 100%. One of the goals of this study was to draw up a set of rules for the performance of triage based on the question type being triaged, which could be utilized as the basis for designing and building a system to automate part or all of the triage process. It is these very strong correlations, presented in Table 4-21, that may be utilized as the basis for designing systems to automate the triage process. Such a system would need to involve two parts: 1) functionality to automatically classify questions as they are received by a digital reference service, and 2) functionality to make a recommendation, based on the classification of the question, of the appropriate triage recipient. This proposed system is discussed further in chapter 5.

4.5. Chapter Summary

This chapter discussed in detail the findings of this study. This study determined 1) What attributes of questions affect the triage process, and 2) How question type correlates with the action taken on a question in the triage process. Expanding on the fifteen factors that Pomerantz, Nicholson, and Lankes (2003) discovered that influence the triage process,

this study discovered eight attributes of questions, and a total of thirty-eight criteria in eight categories, that affect the triage process. Intercoder reliability statistic Cohen's κ was computed between the coders' classifications, and the values of κ ranged from 0.61 to 1, indicating moderate to perfect reliability. The correlation between question type and the action taken on a question in the triage process was determined by calculating Cramér's V, and the values for V ranged from 0.15 to 0.71, indicating some weak to moderately strong correlations when the correlations are considered for entire question taxonomies. Some very strong correlations occur between specific intersections of question types and triage action. It is these very strong correlations that may serve as the basis for designing systems to automate the triage process.