Using large multi-purpose corpora for specific research questions: discourse phenomena related to *wh*-questions in the Spoken Dutch Corpus

Nelleke Oostdijk and Lou Boves
Dept. of Language and Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
{N.Oostdijk, L.Boves}@let.kun.nl

Abstract

In this paper, we investigate whether a dataset derived from a multi-purpose corpus such as the Spoken Dutch Corpus may be considered appropriate for developing a taxonomy of *wh*-questions, and a model of the way in which these questions are integrated in spoken discourse. We compare the results obtained from the Spoken Dutch Corpus with a similar analysis of a large random collection of FAQs from the internet. We find substantial differences between the questions in spoken discourse and FAQs. Therefore, it may not be trivial to use a general purpose corpus as a starting point for developing models for human-computer interaction.

1. Introduction

Over the past years the issue of data sparseness has received a great deal of attention. The almost insatiable need for data arose from the field of linguistic engineering as much as from the field of speech technology. Thus, a great many projects were initiated that were directed at compiling large collections of data. Since the compilation of corpora throughout the years has continued to be costly, both in terms of time investment and manpower, the creation of multi-purpose corpora has prevailed. Examples of corpora that have resulted from this approach are the British National Corpus (BNC; Aston & Burnard 1998) and the American National Corpus (ANC; http://americannationalcorpus.org/), while also the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN; Oostdijk 2000) – although much smaller in size – fits this description. As in recent years it has been the availability and more specifically the quantity of data that has been at the centre of attention, the question whether the data are appropriate for the specific purposes for which they are employed has often been neglected. In the present paper, we report the results of a case study that was conducted and which aimed to address this issue. In this study it was investigated whether a dataset derived from a large multi-purpose corpus such as the Spoken Dutch Corpus may be considered appropriate for developing a taxonomy of wh-questions and a model of their discourse structure. Since the final goal of our work is the development of interactive spoken question answering, we compare the results with an analysis of a corpus of questions obtained from a number of FAQ websites.

2. Background and motivation

Before we could even begin to try and answer the question raised above, it was clear that we needed to establish what kinds of question one would expect to find in the context of spoken QA. Unfortunately, there was no ready answer to this question. Given the present state of the art in the field of question-answering, current QA systems generally are directed towards handling written language input. Moreover, when we consider the nature of the questions that are being asked, it appears that they are mostly stand-alone factoid questions asking about who, what, where, when and how. One of the more recent developments here is that systems are required to handle series of related questions, where answers given by the system trigger follow-up questions, or where the system asks clarification questions before even attempting to find answers. Knowledge acquired in dealing with previous questions must then be employed in order to handle the current query. For the development of future interactive QA one might greatly benefit from a better understanding of how questions are embedded in spoken discourse.

The question we address in this paper is whether a general purpose corpus such as CGN can be used for this highly specific type of discourse analysis. At the same time we aim at making an inventory of the spoken language phenomena that distinguish questions in spontaneous dialogues and conversations from those in interactions by means of a keyboard and screen. In this paper, we limit ourselves to the analysis of *wh*-questions.

3. Data collections

For the present investigation two data collections were used, viz. a subset of data from the CGN and a set of data collected from the internet. These are described in sections 3.1 and 3.2 respectively.

3.1. Data from the Spoken Dutch Corpus

Since the final release of the CGN was not yet available, we used the sixth intermediate release (CGN R6). From the data available in this release only the northern Dutch data were included in our present investigation. By means of the COREX exploitation software we first extracted all questions. This yielded an initial set of 38,101 instances. Through manual selection this set was reduced to include only *wh*-questions. This set was further delimited by requiring that questions should display the unmarked word order normally found in independent *wh*-questions. Thus included were questions like [1]-[2], while excluding instances like [3].[1]

[1] wat betekent botanisch precies?
[2] (nu kun je zeggen van) hoeveel zouden d'r waarschijnlijk zitten aan Serven?

[1] In order to maximize the number of questions that could be obtained and also to allow us to study how in spoken discourse questions are embedded, we decided to include not only independent questions but also dependent ones. However, only those dependent questions were included that answered to the word order criterion.

[3] (weet u) hoe ze klinkt?

Other *wh*-questionlike instances that were excluded are for example [4] and [5].

[4] hij verkocht een wat?
[5] Pisa ligt in wat voor een kleur?

Thus we obtained a dataset comprising 10,033 questions. Table 1 displays the composition of the subcorpus that we used, and the number and distribution of the questions derived.

Component:	**No. of words**	**No. of *wh*-questions**
1. face-to face convers.	1,206,054	5,038
2. interviews	249,787	256
3. telephone conversations	312,279	1,206
4. business negotiations	136,358	327
5. interviews and discuss.	481,509	1,498
6. discuss., debates, etc.	221,531	30
7. lectures	31,569	244
8. descriptions of pictures	0	0
9. spontaneous comment.	84,727	66
10. newsreports, current affairs programmes	35,246	93
11. news bulletins	280,103	0
12. commentary	27,881	25
13. lectures, speeches	67,370	62
14. read aloud text	551,595	1,188
Total	**3,686,009**	**10,033**

Table 1. Number and distribution of *wh*-questions in the CGN (R6 NL)

3.2. Internet data

In order to allow for a comparison between the naturally spoken questions encountered in the CGN and questions one would expect to find in an IR/IE context, we decided to compile a second dataset. We collected data from the FAQs sections of a random set of 104 internet sites representing a variety of domains. From an initial exploration it became apparent that a number of FAQs were in fact not questions in the form of an interrogative structure at all. On a total of 3,709 entries, 99 were declaratives. Characteristically, one would expect to find these in reply to a question prompt like: *How may I help you?* Or *What seems to be problem?* (cf. exs. [6]-[7]).

[6] Ik wil graag meer informatie over duurzame energie.
[7] Ik heb een nieuw e-mailadres, maar ben mijn wachtwoord kwijt.

Another 331 instances were entries where a statement was immediately followed by an explicit question, either a *wh*-question (260 instances; eg ex. [8]) or a polar question (71 instances; eg ex. [9]).

[8] Ik zoek meer informatie over NEN en/of ISO-normen. Waar kan ik die vinden?
[9] Ik heb een draadloze telefoon. Kan ik het bereik daarvan vergroten?

The remaining entries were single questions, 2,324 of which were *wh*-questions, 955 polar.

4. Data analysis

An analysis of the two datasets was undertaken in which we investigated the frequency and distribution of various types of *wh*-question, the effect of the type of speech on the distribution of question types, and the occurrence of reduced questions. In addition, some phenomena were investigated which are commonly associated with spoken discourse. The results of this analysis are presented below.

4.1. Frequency and distribution

The prototypical *wh*-question as described in the literature (eg Donaldson 1997; Haeseryn et al. 1997; de Vries 2001) is introduced by a *wh*-element.[2] The *wh*-element is either an interrogative pronoun or an interrogative adverb. In Dutch, *wh*-elements (pronouns and adverbs) take on a variety of forms. We decided to classify these in 7 major types: hoe (*how*), waar (*where*), waarom (*why*), wanneer (*when*), wat (*what*), welk(e) (*which*), and wie (*who*). Figure 1 gives of an overview.

hoe: *hoe, hoeveel, hoeveelste, hoevelen, hoever, hoeverre, hoezeer, hoezo*.
waar: *waar, waaraan, waarbij, waarbinnen, waardoor, waarheen, waarin, waarlangs, waarnaar, waarnaartoe, waaromheen, waarop, waarover, waartegen, waaruit, waarvan, waarvandaan, waarvoor, waarzo*
waarom: *waarom*
wanneer: *wanneer*
wat: *wat, wat voor, wat voor een, wablief, watte, wattes, gewat*
welk(e): *welk, welke*
wie: *wie, wie d'r, wiens, wier*

Figure 1. Types of *wh*-question

The distribution of question types is roughly the same for the two datasets (cf. Table 2) and across various types of speech: wat and hoe are the most frequent types of question (cf. Table 3).[3] Compound questions involving more than one *wh*-element were classified as multiple.

Type of *wh*-question	**Absolute freq**		**Relative freq.**	
	CGN	**WWW**	**CGN**	**WWW**
hoe	2,792	1,014	27.83	39.24
waar	985	196	9.82	7.59
waarom	972	201	9.69	7.78
wanneer	307	80	3.06	3.10
wat	3,797	811	37.85	31.39
welk(e)	411	185	4.10	7.16
wie	672	51	6.70	1.97
multiple	97	46	0.97	1.78
total	**10,033**	**2,584**	**100.00**	**100.00**

Table 2. Frequency and distribution of question types across the two datasets

[2] More accurately, a *wh*-element or a constituent (usually a prepositional phrase) containing such an element. Deviant structures may occur as a result of, for example, topicalisation or the initial placement of conditional clauses.

[3] Table 3 gives the relative frequencies for the four major text types (components 1, 3, 5 and 14 in the CGN dataset respectively; cf. Table 1).

Type of *wh*-question	Relative freq.			
	CGN-1	CGN-3	CGN-5	CGN-14
hoe	28.78	33.00	27.10	22.90
waar	9.77	10.45	8.68	12.21
waarom	7.80	7.88	11.75	17.26
wanneer	3.28	4.98	2.47	2.19
wat	37.44	35.49	39.52	35.44
welk(e)	5.00	2.74	3.94	1.94
wie	7.15	4.64	5.47	7.66
multiple	0.79	0.83	1.07	0.42
total	**100.00**	**100.00**	**100.00**	**100.00**

Table 3. Frequency and distribution of question types across different types of speech

For all 7 classes it holds true that the central *wh*-element is the most frequent by far. Thus, in the spoken data, in 2,247 out of 2,792 questions in the hoe-class, the *wh*-element is realized by *hoe*; for waar, wat and wie the respective figures are 596/985, 3518/3,797 and 668/672. In the internet data, all but 91 questions can be accounted for by the central *wh*-elements, 64 of these contain the wh-word *hoeveel*.

4.2. Reduced questions

What goes undetected in the presentation of the frequency and distribution information as presented in Tables 2 and 3 is the role played by reduced *wh*-questions, ie questions in which essentially only the *wh*-element remains, while the verb and possibly other constituents are omitted. Examples are [10]-[12].

[10] in welk opzicht?
[11] sinds wanneer?
[12] naar wat voor quiz?

Such questions typically serve the purpose of obtaining clarification from the interlocutor for something that was introduced earlier on in the discourse. It is therefore no surprise that reduced questions occur most frequently in the more interactive text types in the spoken data. Thus, on a total number of 10,033 questions, 1,450 (14.45%) are reduced. The proportion of reduced questions in face-to-face conversations and telephone conversations is 18.12% and 18.91% resp. It is also with these two types of speech that there is the frequent use of (reduced) 'repeat' questions: questions that prompt the interlocutor to repeat (specific bits of) what he/she said. Examples are [13] and [14].[4]

[13] ja hoe oud?
[14] wauw hoeveel?

Table 4 displays the frequency and distribution of reduced questions across different question types in the two types of conversational speech (CGN-1 and CGN-3).

As is apparent from Table 4 especially waarom and welk(e) questions are particularly prone to reduction.[5]

Type of *wh*-	CGN-1		CGN-3	
	no.	%	no.	%
hoe	196	13.52	35	8.79
waar	63	12.80	17	13.49
waarom	**164**	**41.73**	**65**	**68.42**
wanneer	24	14.55	6	10.00
wat	307	16.28	73	17.06
welk(e)	**75**	**29.76**	**17**	**51.52**
wie	79	21.94	15	26.79
multiple	5	12.50	0	0
total	**913**	**18.12**	**228**	**18.91**

Table 4. Frequency and distribution of reduced questions across different types of speech

By comparison, reduced questions in the internet data are very few and show virtually no variation. Only 42 instances were encountered, 38 were questions following a statement. All 28 reduced wat questions are of the form *Wat nu?* The two reduced hoe-questions are realized as *Hoe verder*? and *Hoe nu verder?* The hoe- and wat-questions without exception carry the meaning 'how should I proceed?' Finally, the remaining 12 questions were waarom-questions asking for clarification.

4.3. Introductory elements

In naturally spoken language, *wh*-questions are commonly introduced by one or more introductory elements that precede the *wh*-element. Figure 2 gives an overview of the main categories and the elements they comprise. [6]

Connectives: *dus, en, enne, maar, of, van, want, …*
Hesitations: *uh, uhm*
Initiator: *alleen, goed, hé, kijk, nou, trouwens, tja, zeg, …,*
Reaction signals: *ach, ah, goh, hè, ja, nee, oh, oh ja, oh nee, mmm, nee, oké, precies, …*
Vocatives

Figure 2. Introductory elements

In all, 3,906 questions in the CGN contain one or more introductory elements. Table 5 lists the frequencies of the 10 most frequent single introductory items, which together account for 1,731 instances. [7]

item	frequency	item	frequency
en	664	*want*	54
maar	411	*oh*	53
ja	236	*of*	40
uh/uhm	144	*hé*	34
nou	66	*nee*	29

Table 5. The 10 most frequent introductory elements in the CGN dataset

Noteworthy is that in the internet data no introductory elements were encountered.

[4] In this context it is worth mentioning that some full questions are in fact formulaic and can serve the same purpose. These include *wat/hoe zeg/zei je?* and *wat/hoe zegt/zei u?* In informal conversation they are commonly reduced to *wat?* and *hoe?*

[5] Interestingly. a very large proportion of reduced hoe questions (viz. 204 out of 297 or 68.69%) contains the *wh*-element *hoezo* (*hoezo, hoezo dan, hoezo niet*) which is roughly equivalent to *how's that* or *why*.

[6] Dependent *wh*-questions are commonly embedded by means of the connective *van* or *of*, or a reporting clause.

[7] Not included here are the vocatives that account for 37 instances.

4.4. Answer prompts

Another phenomenon found in our material exclusively with the naturally spoken data is that questions may already contain a candidate answer. Consider exs. [15]-17].

[15] wat uh Van Dale?
[16] en waar komt ie vandaan Leiden Amsterdam?
[17] of ja wat zei ik net De Nederlandse Bank?

Such questions characteristically are used with the intention of verifying information. Judging from our data, their frequency is extremely low.

4.5. Reference

As mentioned before, in the internet data two groups of questions can be distinguished: the full independent questions that occur by themselves, and the sometimes reduced questions that follow a statement. Questions in the first group are usually self-contained. The questions in the second group, however, almost invariably require anaphora resolution (cf. exs. [18]-[19]).

[18] Ik kan mijn password niet wijzigen. Hoe kan dit?
[19] Reizigers in besmette gebieden worden gescreend op SARS. Wat is dat?

In the spoken data, the picture is rather different. A large proportion of full independent questions require anaphora resolution. However, with our present dataset it is impossible to determine whether the referent is to be found in a preceding utterance produced by the present speaker or in an utterance produced by the interlocutor.

With regard to the matter of reference, two further observations can be reported. One relates to the phenomenon of topicalisation (cf. ex. [20]). Although topicalisation is generally assumed to be characteristic of spoken rather than written language, the number of topicalised questions in the spoken data is negligible: in all we counted 91 instances. The other observation concerns the use of cataphoric reference (cf. exs. [21]-[22]). This, again, is a phenomenon found in our datasets exclusively in the spoken data and even then, it occurs very rarely.

[20] en sinussen wat zijn dat precies voor dingen?
[21] waardoor wordt dat veroorzaakt die temperatuur-stijging?
[22] waar ligt dat dan Oorschot?

5. Discussion and conclusion

A comparison of the two datasets has brought to light a number of differences – quantatively and qualitatively – that may be considered relevant when contemplating the issue of whether a dataset derived from a large multi-purpose corpus such as the Spoken Dutch Corpus may be considered appropriate for developing an NLP system that can support natural interaction in a spoken QA system.

Our analysis leads us to conclude that in principle CGN data are appropriate for developing a model of the (*wh*-)questions that people will use in interactive QA. CGN constitutes a very rich source both in terms of the number of questions as well as in terms of the (structural) variation that is encountered. All possible (likely) variants are represented in the data and the coverage of a language model that has been developed on the basis of these data will appear to be adequate for handling all sorts of spoken *wh*-questions. The data also make it possible to model a number of phenomena that appear to be characteristic of naturally spoken questions. Here it is useful to distinguish phenomena that are unique in human-human interaction (eg vocatives, formulaic questions, reaction signals, expletives), while other phenomena are found in both human-human and human-machine interaction (eg connectives, hesitations, false starts). In deriving a dataset for modeling questions in spoken QA it seems a good idea to create the equivalent of a stop list for formulaic questions (*wat zeg je?, hoe bedoel je?, hoezo? hoe gaat het?*).

However, we should add a word of caution. As was pointed out before, the present study has been limited to *wh*-questions. On the basis of what we have seen in the internet data, it would appear that this limitation can not be upheld. Imposing on people that they should use only *wh*-questions might prove to be too severe a limitation to a QA system that is intended to handle natural interaction. Therefore, further research is necessary into alternative ways of asking for information that the user is likely to use. Other issues that future research should address include the following: Can we distinguish between (the types of) questions that are likely to be asked by the user of a QA system, and questions that the system may generate? And related to that How can we distinguish between a starter question and follow-up questions?

In the present paper, deriving a dataset from the CGN was said to be a case study. Before we can generalize the findings of the present research to other corpora and other languages, similar studies should be undertaken with aim of corroborating the present findings.

6. Acknowledgement

Thanks are due to Johan de Veth who raised the question addressed here and thus put us onto the research reported on in this paper.

This publication was supported by the Netherlands Organization for Scientific Research (NWO) under grant number 014-17-510.

7. References

Aston, G. and L. Burnard, 1998. *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Corpus Gesproken Nederlands (Spoken Dutch Corpus). *Release 6, November 2002. Annotaties/Annotations. CGN_R6_ANN1*. Distributed by ELDA, Paris.

Donaldson, B. 1997. *Dutch. A Comprehensive Grammar*. London, New York: Routlegde.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij, and M.C. van den Toorn, 1997. *Algemene Nederlandse Spraakkunst*. Groningen: Martinus Nijhoff.

Oostdijk, N. 2000. The Spoken Dutch Corpus. Outline and first evaluation. In: M. Gravilidou, G. Caravannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC). 31 May-2 June 2000*. Athens, Greece. Vol. 2: 887-894.

TREC. http://trec.nist.gov/

Vries, de J. 2001. *Onze Nederlandse Spreektaal*. Den Haag: SDU Uitgevers.