Chapter 5

A corpus study

1 Introduction

The experimental results in the previous chapters showed that the resolution of subject-object order ambiguities in Dutch is affected by the nature of the NPs involved. In this chapter, word order preferences will be investigated using a different method, namely by comparing the frequencies of occurrence of subject-object and object-subject clauses in a sample of Dutch texts. The first aim of this study is to see which order of subject and object is the most frequent in naturally produced wh-questions, and whether the relative frequency of the two orders is different depending on the type of the NPs. Frequency data may therefore provide additional support for the view that order preferences are determined by various factors in addition to the syntactic bias for a subject-object order.

The second aim of this study is to see to what extent frequency data correspond to comprehension data. The relation between frequency and parsing preferences is important in the light of current theories on sentence processing. Some theorists claim that the sentence processing mechanism is mainly frequency-driven: in the case of structural ambiguity, the parser opts for the analysis that is the most frequent in the language. Frequency-based theories thus predict a close correspondence between frequency and processing difficulty: structures that are relatively infrequent are also the ones that cause an increase in processing difficulty. For previous studies comparing corpus and experimental data, see Cuetos, Mitchell and Corley (in press); Gibson and Pearlmutter (1994), Gibson, Schütze and Salomon (1996), Hindle and Rooth (1993), Spivey-Knowlton and Sedivy (1995), among others.

One problem in comparing experimental and frequency data is the problem of grain-size (cf. Mitchell, Cuetos, Corley and Brysbaert, 1995): which criteria are relevant in classifying utterances and their frequencies? For instance, the frequency of the subject-object and object-subject order may be determined separately for wh-questions and other clause types, e.g. declaratives. However, at some level of generalization, wh-questions may be lumped together with other clause types. The frequency of subject-object and object-subject clauses may then not be available for wh-questions separately, but pooled over various clause types. Alternatively, finer-grained distinctions are conceivable: separate subtypes of wh-questions may be distinguished on the basis of e.g. the nature of the verbs and NPs involved. For each of these subclasses the number of subject-object and object-subject clauses then has a separate influence on processing. As we will see below, the relative frequency of the alternative readings may differ depending on
the grain-size. Whether or not experimental and frequency data converge may thus rely on the grain-size assumed by the researcher.

Frequency-based theories differ with respect to which information the sentence processor is assumed to be sensitive to in analyzing utterances and storing frequency information (cf. Mitchell, Cuetos, Corley and Brysbaert, 1995). According to the linguistic tuning hypothesis proposed by Mitchell (Mitchell, 1994; Mitchell et al., 1995), only coarse-grained generalizations play a role in on-line processing. These are generalizations that abstract away from the properties of the verbs and NPs involved. Other theories assume that fine-grained (lexical) information plays a role in parsing (e.g. Spivey-Knowlton and Sedivy, 1995), or that generalizations at various levels of abstraction are relevant (e.g. MacDonald, Pearlman and Seidenberg, 1994). The present study may provide some more insight into this grain-size problem.

Below the frequency of occurrence of subject-object and object-subject clauses in Dutch is investigated. For practical reasons, the counts are restricted to welke ‘which’-questions. In Section 2, I will first give a description of the corpus, and the coding and counting procedures. Results will be reported in Section 2.3. In Section 3, the frequency data will be compared to the experimental data reported in the previous chapters, and the grain-size problem will be discussed.

2 Order frequencies in welke-questions

2.1 Introduction

The relative frequencies of subject-object versus object-subject welke-questions in Dutch were estimated on the basis of a representative sample of naturally occurring written texts. From this text corpus, all welke-questions were extracted, and the numbers of subject- (object) and object-subject welke-clauses were counted. In order to compare the frequency data to the experimental data and to investigate which grain-size would be the most appropriate, the subject-object to object-subject ratio was computed as a function of several factors.

First, subject- and object-initial clauses were counted separately for main and embedded clauses. Main and embedded clauses differ in the position of the finite verb: the verb appears clause-initially or after the first constituent in main clauses, but in final position in embedded clauses. This may have consequences for the relative frequency of the alternative orders.

Second, the number of NPs in the clause was assumed to be relevant. In the experiments to date, only clauses have been used that contained two argument NPs: one functioning as the subject, one as the direct object. Frequency data for clauses containing (at least) two argument NPs may therefore more closely resemble the experimental data than when the number of NPs is not taken into consideration. The number of subject-object and object-subject clauses was therefore determined twice: once on the basis of all welke-questions in the corpus,
irrespective of the number of NPs in the clause; and once for only those clauses that contained at least a subject and an object NP.

Third, in order to investigate the effect of the nature of the second NP, the frequencies of the alternative orders were investigated separately for (di)transitive clauses containing (i) a pronoun, (ii) a full definite NP, or (iii) an indefinite NP in the second NP position. In Chapter 4 we have seen that the type of the second NP plays a role in ambiguity resolution: when the second NP was a pronoun, an object-subject preference was attested (Experiments 3-6), whereas a tendency for a subject-object preference was seen for definite NPs (Experiment 7). If there is a close correspondence between comprehension and frequency, one would expect the object-subject order to dominate in clauses containing a pronoun; the subject-object order is expected to be the most frequent in clauses containing a full definite NP as the second NP.

Finally, a distinction was made between NPs referring to animate entities and NPs referring to inanimate entities. The frequency of the subject-object versus object-subject readings is likely to depend on the animacy properties of the NPs (cf. MacWhinney, Bates and Kliegl, 1984; McDonald, 1987). Furthermore, in the experiments conducted in the previous chapters, both NPs always referred to animate entities. Restricting the corpus to cases in which the NP referents do not differ in their animacy properties, or to cases in which both NPs refer to animate entities only, may thus be more informative in comparing frequency and experimental data.

2.2 Methods

2.2.1 The corpus

The search was carried out using one of the electronically accessible Dutch corpora provided by the Instituut voor Nederlandse Lexicologie, Leiden. The corpus used in the present study covers 5 million words, and consists of passages drawn from various sources of written materials: newspapers, magazines (on environment, linguistics, politics and leisure), books (on environment, linguistics, business and employment, and politics) and texts that were written to be spoken (news broadcasts and political speeches).

2.2.2 Restricting the corpus

The corpus contained 1362 utterances in which the word welke or the neuter form welk occurred.¹ To allow a comparison with the experimental data, the set was narrowed down to those utterances in which the welke-phrase was the first phrase of a finite question and was either subject or object. Such clauses are structurally

¹ In the following I will use welke as a cover term for both welke and welk.
similar to the wh-clauses used in the experimental materials. The following restrictions were therefore made to obtain the relevant subset.

First, clauses in which the welke-phrase was part of a PP were excluded by extracting only those questions in which the welke-phrase was not preceded by a preposition. This yielded a smaller set of 792 utterances. Unfortunately, this also excluded welke-clauses introduced by e.g. discussiëren over (‘to discuss’), het besluit over (‘the decision about’) which might have been relevant for our purposes. Omission of these items would probably not have qualitatively affected the pattern of results obtained, as there is no straightforward reason why welke-questions introduced by such nouns and verbs should behave differently from other welke-questions.

The set was further restricted by excluding the following cases: (i) clauses in which welke introduced a relative clause (72 cases); (ii) clauses in which welke was used in expressions such as welke dan ook ‘no matter which’ (30 cases); (iii) clauses in which welke was used as an exclamatory expression as in welk groot musicus ons is ontvallen ‘what a great musician we lost’ (four cases); (iv) clauses in which the welke phrase did not appear in clause-initial position (six cases); (v) infinite clauses, e.g. welke te kiezen ‘which one to chose’ (two cases); (vi) clauses in which the welke phrase clearly was the predicate of a copula construction, as indicated by verb agreement: zeg welk schip je bent ‘say which ship you are’ (one case); (iv) ellipsis and gapping constructions (51 cases). In e.g. Een andere aanpak is dus nodig, maar welke? ‘A different approach is therefore needed, but which one?’ the wh-phrase was not counted; in welke zaken hij wel en welke zaken hij niet zelf zou willen behandelen ‘which affairs he did and which affairs he did not like to deal with himself’ only one wh-phrase was counted. Furthermore, (vii) a conjunction of two welke-phrases was treated as one phrase (three cases). Finally, (vii), sixteen cases were dropped because they were a literal repetition of a sentence already counted. For instance, in one of the passages, the sentence Welke coalitie zou na de volgende verkiezingen jouw voorkeur hebben? ‘Which coalition would you prefer after the next elections?’ was used over and over again for rhetorical reasons, but only counted once.

Thus a total of 607 clauses was obtained which started with welke or welk functioning as the subject or object of the clause.

2.2.3 Coding

The welke-clauses thus obtained were coded to indicate (i) whether the wh-phrase was subject, object or indirect object; (ii) whether the clause was main or embedded; (iii) whether the predicate contained one or more than one argument NP; (iv) whether the second NP, if available, was a pronoun, definite or indefinite NP; and (v) whether the wh-phrase and the second NP, if available, referred to an animate or inanimate entity. Utterances were parsed by hand and coded independently by the author and a second judge. Cases each had initially coded differently were decided upon by discussion.
2.2.3.1 Subject / object-initiality
The welke-phrase was coded as the subject of the clause on the basis of number information at the verb or case marking of the second NP. If the clause was not syntactically disambiguated by these means, the grammatical function of the welke-phrase was determined on the basis of semantic information.

Twenty-nine cases were somewhat problematic. These were copula constructions containing a second NP, such as welke dat is ‘which one that is’. Here, the wh-phrase can be interpreted either as the subject of the clause or as the predicate of the copula construction. In the results section below, two figures will be given, one in which the wh-phrase in such constructions is interpreted as the subject, and one in which such copula constructions are left out.

2.2.3.2 Main and embedded clauses
Main and embedded welke-clauses were primarily distinguished on the basis of the position of the finite verb: the finite verb appeared clause-finally in embedded clauses, and directly followed the welke-phrase in main clauses. In a few cases, the second and clause-final position coincided. The clause was then coded as an embedded clause if it fulfilled a role in a hierarchically higher clause, for instance, if it was the subject of a clause, the complement of a noun or a verb, or part of a copula construction. In all remaining cases, the welke-clause was coded as a main clause.

2.2.3.3 The number of argument NPs
A distinction was made between clauses containing only one argument NP (intransitives), on the one hand, and clauses containing at least a subject and an object NP ((di)transitives) on the other. The class of intransitive predicates consisted of (i) real intransitive predicates, e.g. ‘to dream’ occurring without a direct object; (ii) ergatives e.g. ‘to disappear’; (iii) passives, e.g. ‘to be chosen’; (iv) copula constructions e.g. ‘to be necessary’; (v) cases in which the verb selected a sentential complement; and (vi) reflexive predicates: most (five cases) were inherent reflexives such as zich voordoen ‘to happen’; the remaining two cases were zich melden ‘to report’ and zich NP noemen ‘to call oneself NP’.

A few cases deserve special mention. First, in five cases the verb vinden or achten (‘to consider’) did not have a sentential complement as its object, but an NP and a secondary predicate. An example is: Welke Nederlander vindt ook het omgekeerde de normaalste zaak van de wereld ‘Which Dutchman also considers the reverse the most common thing in the world’. These cases were treated as if the second NP (het omgekeerde) was the direct object of the verb, and thus were coded as transitive predicates.

Second, the corpus contained six instances of ditransitive predicates. In all cases, the welke-phrase was either the subject or the direct object of the clause. In only one case, the second NP was the indirect object (welke exploitant de PTT concurrentie mag aandoen ‘which operator is allowed to offer competition to the PTT.’) This case was coded as a transitive, subject-object clause.
Third, the corpus included ten standard expressions. Since these were not fully idiomatic (for instance, question formation did not substantially change the meaning) we chose to include them in the final sample. The clauses containing the expressions *kans maken op* ‘to stand a chance of’ (two occurrences), *aanleiding geven tot* ‘to provoke’ and *recht doen aan* ‘to do justice to’ (each occurring once) were coded as transitives, with the indefinite NP as the object. Furthermore, the one instance of *het voor elkaar krijgen* ‘to manage to do it’ was coded a transitive with the pronoun as the object. Finally, the clause *welke gedeelten het hier betrof* ‘which parts it concerned here’, and the four instances of *welke kant NP opgaat* ‘which way NP goes’ were counted as object-subject transitives.

As the number of these special cases is fairly small, inclusion of these cases does not substantially affect the pattern of results reported below.

2.2.3.4 The type of second NP
If available, the second NP was coded as one of the following classes: (i) pronoun: personal or demonstrative pronouns (*die*, *dat* ‘that’, *deze*, *dit* ‘this’); (ii) definite NP: proper names, and NPs starting with the definite article *de*/*het* ‘the’, a demonstrative article *die*, *dat* ‘that’, *deze*, *dit* ‘this’ or *zulk(e)* ‘such a’, a possessive pronoun, or the quantifiers *beide* ‘both’ and *elke* ‘every’; and (iii) indefinite NP: bare plurals and singulars, and singular NPs starting with the indefinite article *een* ‘a’.

2.2.3.5 Animacy
*Welke*-phrases and second NPs referring to people or animals were considered animate. This class also included NPs denoting political parties, firms, institutions and other groups of people. All other noun phrases were coded as inanimate.

2.3 Results

2.3.1 Frequency of S(O) versus OS structures

The frequency of occurrence for subject-initial and object-initial *welke* questions for all predicates, intransitives and (di)transitives, is given in Table 5.1. In a significant majority of cases, 64%, the *welke* phrase was the subject of the clause \[ \chi^2 (1) = 48.2, \ p < .001 \]. The corpus contained fewer main than embedded clause questions. However, main and embedded clauses did not differ in the relative frequency of the two orders: 61% of the main and 65% of the embedded clauses were subject-initial, a non-significant difference \[ \chi^2 (1) = .519, \]
In Table 5.1, copula constructions with a second NP such as welke dat is ‘who that is’ were counted as subject-initial clauses (cf. Section 2.2.3.1). When clauses containing such predicates were dropped, the proportion of object- and subject-initial clauses did not change substantially: 78 out of 132 main clauses, and 282 out of 446 embedded clauses were subject-initial, yielding a total of 360 (62%) subject-initial versus 218 (38%) object-initial clauses. This difference remained significant $\chi^2 (1) = 34.1, p < .001$. Again, main and embedded clauses did not differ $\chi^2 (1) = .577, p = .447$.

In more than half (55%) of the cases considered in Table 5.1, the welke-phrase was the only argument NP in the clause. In these cases, the wh-phrase was the subject just because it was the only NP available. The numbers in Table 5.1 thus do not provide any information concerning the order of subject and object when both are expressed. It is therefore more informative to investigate the frequencies of the subject-object and object-subject orders for clauses containing two or more NPs. A separate analysis was therefore carried out using a restricted set containing (di)transitive predicates only. Results are given in Table 5.2.

When counts were restricted to clauses containing at least two NPs, subject-initial clauses were less frequent than object-initial clauses. In total, only 21% of the (di)transitive welke-clauses were subject-initial, versus 79% object-initial. This was significantly different from an equal distribution of subject and object-initial clauses $\chi^2 (1) = 94.3, p < .001$. Again, there was hardly any

$\chi^2$ refers to the chi-squared statistic, a measure of the difference between observed and expected frequencies. The notation $\chi^2 (1)$ indicates that the chi-squared statistic is calculated with one degree of freedom, which is the case when comparing two groups.

---

2 Yates’ continuity correction is applied in this and following tests comparing two groups (cf. Hays 1988).
difference between main and embedded clauses: 25% of the main and 19% of the embedded clauses are subject-object. This difference was not significant \[ \chi^2 (1) = .76, p = .383 \].

To sum up, no differences were attested between main and embedded welke-questions. However, the number of NP arguments had a large effect on the relative frequency of subject- versus object-initial order of the welke-clause. Collapsing over all types of predicates, subject-initial clauses were more frequent than object-initial ones; however, if only (di)transitive predicates are taken into consideration, object-initial clauses were the most frequent.3

2.3.2 Order as a function of the second NP

Now let us see whether the nature of the second noun phrase makes a difference with respect to the relative frequency of the subject-object and object-subject orders. In Table 5.3, the order frequencies are split for clauses containing a pronoun, a definite noun phrase, or indefinite NP as a second noun phrase. Due to the small number of items in some cells, I will no longer distinguish main and embedded clauses. From the results reported above there is little reason to expect these structures to differ.

In all conditions subject-object clauses were less frequent than object-subject clauses. The nature of the second NP had a significant effect on the

---

3 The omission of relative clauses starting with welke is unlikely to have substantially affected the results: 58 out of the 72 relative clauses (81%) were subject-initial. When the counts were restricted to (di)transitives, 8 out of 22 (36%) clauses were subject-initial. These figures pattern with the question data.
relative frequency of this reading $\chi^2 (2) = 34.47, p < .0001$. When the second
NP was a pronoun, the wh-phrase was the subject of the clause in only 6% of the
cases. The proportion of subject-object clauses became somewhat larger when the
second noun phrase was a definite NP (31%) and was largest when the second NP
was an indefinite NP (40%).

The difference between clauses containing a pronoun and clauses
containing a definite second NP was significant $\chi^2 (1) = 25.27, p < .0001$, as
was the difference between pronouns and indefinites $\chi^2 (1) = 27.93, p < .0001$.
Definite and indefinite NPs, however, did not differ $\chi^2 (1) = .57, p = .450$.
Within-class comparisons showed that the number of object-subject clauses and
the number of subject-object clauses was significantly different if the second NP
was a pronoun $\chi^2 (1) = 100.54, p < .001$ or a definite NP $\chi^2 (1) = 14.49,
 p < .001$, but not if the second NP was indefinite $\chi^2 (1) = 1.88, p = .170$.

In sum, across-the-board, the object-subject reading for transitive
teradicats is the most frequent one. The definiteness of the second NP has an
effect: object-subject clauses occur relatively more often if the second NP is a
pronoun than if it is a definite or indefinite NP.

2.3.3 Controlling for animacy differences

The figures above should be interpreted with caution, however. The large number
of object-subject clauses may be related to the semantic properties of the welke-
phrase and the second NP. In Table 5.4 the numbers of subject-object clauses are
displayed as a function of the animacy properties of the welke-phrase (NP1) and
the second NP (NP2). When the two NPs differed in their animacy properties, the
NP denoting the animate entity was the subject in most of the cases: when the
first NP was animate and the second was inanimate, the order was subject-object
in 95% of the cases; when the first NP was inanimate and the second NP animate, the order was subject-object only in 3% of the cases, which implies that in 97%, the second, animate NP was the subject.

Note that the welke-phrase denoted an inanimate entity and the second NP an animate one in more than half of the cases in the sample. This may have affected the frequency of the object-subject order as given in Table 5.3. Recall that the experimental materials only contained NPs that both referred to animate entities. Since one of the goals is to see to what extent the frequency data correspond to the experimental data, it may be more informative to look at the effect of the second NP for only those clauses in which the two NPs do not differ in their animacy properties. In Table 5.5 the numbers are given for clauses in which the NPs both denote animate or both denote inanimate entities.

When animacy differences were controlled for, object-subject clauses were still more frequent on average than subject-object clauses [χ²(1) = 11.89, p < .001]. The type of the second NP still had a significant effect on the subject-object to object-subject ratio [χ²(2) = 13.40, p < .001]. Only 8% of the clauses containing a pronoun, and 35% of the clauses with a definite second NP was subject-object. For clauses containing an indefinite NP as the second NP, the subject-object order became the most frequent order (57%). Pairwise tests showed that pronouns and definite NPs have a different effect on order [χ²(1) = 5.51, p < .025], as do pronouns and indefinite NPs [χ²(1) = 11.50, p < .001]. The difference between definite and indefinite NPs was still not significant [χ²(1) = 2.21, p = .137]. For clauses containing a pronoun or a definite NP as a second NP, but not for clauses containing an indefinite NP, the difference between the two orders was significant [pronouns: χ²(1) = 18.62, p < .001; definite NPs: χ²(1) = 4.74, p < .05; indefinite NPs: χ²(1) = .39, p = .532]. Hence, even when animacy differences between the wh-phrase and the second NP are controlled for, the object-subject order is the most frequent, and is relatively more frequent when the second NP is a pronoun than when it is a definite or indefinite NP.

Note, however, that in the experiments presented in previous chapters the wh-phrase and the second NP both denoted animate entities. If one wishes to
compare frequency and experimental data, it may be necessary to further restrict the set of welke-questions to those containing animate NPs only. Unfortunately, in the present corpus, only 11 utterances meet this criterion. An overview is given in Table 5.6.

**Table 5.6**

*Absolute numbers of subject-object and object-subject welke-clauses as a function of the type of the second NP. NP1 and NP2 both denote animate entities.*

<table>
<thead>
<tr>
<th>type of NP2</th>
<th>order</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>subject-object</td>
<td>object-subject</td>
<td>total</td>
<td></td>
</tr>
<tr>
<td>pronoun</td>
<td>2 (8%)</td>
<td>24 (92%)</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>definite NP</td>
<td>19 (35%)</td>
<td>35 (65%)</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>indefinite NP</td>
<td>13 (57%)</td>
<td>10 (43%)</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>34 (33%)</td>
<td>69 (67%)</td>
<td>103</td>
<td></td>
</tr>
</tbody>
</table>

* This is an instance of a ditransitive predicate. The second NP is the indirect object.

Numerically the set contained more object-subject than subject-object clauses. A $\chi^2$ test revealed no significant differences between the two orders, however
\[ \chi^2 (1) = .818, \ p = .366 \], which is not surprising given the small number of cases. The small sample size does not allow a statistical test of the effect of the second NP. The pattern roughly resembles the one in Table 5.5. There is therefore no reason to suppose that the type of the second NP has a substantially different effect when only animate NPs are taken into consideration.

To summarize, when the welke-phrase and the second NP differ in their animacy properties, the NP referring to the animate entity is more frequently the subject than the object of the clause. However, the relative frequency of the object-subject order was not completely due to differences in animacy: the object-subject order remained the most frequent when animacy differences were controlled for. Furthermore, the effect of the type of the second NP was comparable to the unrestricted set. A comparable pattern of frequencies was also obtained when the set was even further restricted to clauses containing only animate NPs. However, the number of cases in this set was too small to statistically test the differences.

### 2.4 Summary of the findings

The corpus search on welke-questions yielded the following results:

- Collapsing over verb phrase types, the subject-initial welke ‘which’-questions are more frequent than object-initial questions.
- When the sample was restricted to (di)transitive predicates, object-subject welke clauses were more frequent than subject-object clauses;
- Main clauses did not differ from embedded clauses as to the relative frequency of subject-object and object-subject clauses.
- The nature of the second NP matters. Although the object-subject reading was the most frequent, the object-subject order was relatively more frequent when the second NP was a pronoun than when it was a definite or indefinite NP. This difference also holds when the wh-phrase and the second NP both refer to animate or both refer to inanimate entities.

### 3 Discussion

The first aim of this study was to see which order of subject and object was the most frequent for welke-questions in naturally produced texts, and whether the relative frequency of the two orders would differ depending on the type of the second NP. The results show that the subject-initial order is the most frequent if the number of NPs is not taken into consideration. If the sample is restricted to welke-clauses containing at least two NPs, the object-subject order is the most frequent. The relative frequency of the two orders is indeed affected by the type
of the second NP: the number of subject-object clauses was lowest when the second NP was a pronoun, and highest when the second NP was an indefinite NP. The relative frequencies of the two orders differed significantly between clauses containing pronouns on the one hand and clauses containing a definite or indefinite second NP on the other. Hence, also in naturally produced texts, word order is related to the properties of the NPs involved.4

The second aim of this study was to investigate to what extent frequency data correspond to parsing preferences. According to frequency-based approaches

4 Note that in language production, the choice is not only between a subject-object or object-subject order of the wh-clause but also between an active or a passive mode: roughly speaking, if the wh-phrase denotes the agent of the action, the wh-phrase can either be encoded as the subject of an active subject-object clause (ia), or the object of a clause-initial by-phrase in a passive clause (ib); if the wh-phrase is the theme, it either is the object of an active object-subject wh-clause (iia), or the clause-initial subject of a passive (iib).

(i) a. Welke dichter groet hem?
   which poet greets him-ACC
b. Door welke dichter wordt hij gegroet?
   By which poet is he-NOM greeted

(ii) a. Welke dichter groet hij?
   which poet greets he-NOM
b. Welke dichter wordt (door hem) gegroet?
   which poet is (by him-ACC) greeted

If the choice of the subject and object in production were determined solely by the type of the second NP, a second NP pronoun should be more frequently used as the subject than as the object. This means that passives such as (ib) should be more frequent than actives such as (ia). On the other hand one would expect (iia) to be more frequent than (iib) with the pronoun overtly expressed in the by-phrase. The preference for (ib) over (ia), and for (iia) over (iib) should be smaller or reversed if the second NP is a full (in)definite NP. Unfortunately, the corpus was too small to draw any reliable conclusions about these structures. The corpus used contained only two door+wh passives like (ib): one in which the second NP was a pronoun, one with a non-pronominal definite second NP. In contrast, the numbers of active subject-object clauses (ia) were 7 (for second NP pronouns), and 33 (for full definite NPs). These numbers suggest that the active mode is preferred even if the second NP is a pronoun. The frequency data for clauses such as (ii) point into the same direction. Irrespective of the type of second NP, the corpus contained 218 transitive and ditransitive actives like (iia), and 134 passives like (iib). However, the by-phrase was overtly expressed only in eight cases: seven containing a non-pronominal definite NP (versus 72 actives like (iia)); one containing an indefinite NP (26 actives). No cases were found in which the by-phrase contained a pronoun (126 actives). In sum, the sparse data suggest that although the type of second NP may have an effect on the mode, the use of passives and especially passives containing a by-phrase is constrained by additional factors.
to parsing, frequency should correspond to experimental data at least at some level of abstraction. An important question is which grain-size is the most appropriate. I will first compare the corpus and the experimental data for main clauses. Next the effect of the second NP will be dealt with. We will see that the frequency and the experimental data do not fully converge, resulting in an apparent paradox with respect to grain-size. Several solutions will be proposed to resolve the paradox.

3.1 Main clauses and the effects of transitivity

Consider first the order preferences seen in main clause welke-questions. An example of the materials in Experiment 1 is given below:

(1) a. Welke assistenten hadden de professor niet geholpen? [SO]
    which assistants had-pl the professor not helped
    ‘Which assistants didn’t help the professor?’

b. Welke assistenten had de professor niet geholpen? [OS]
    which assistants had-sg the professor not helped
    ‘Which assistants didn’t the professor help?’

These clauses showed a preference for the subject-object order (1a). This corresponds to the frequency data for welke-questions pooled over transitive and intransitive predicates (Table 5.1). However, the experimental data do not correspond to the corpus data when only transitive predicates are taken into consideration: for these cases the object-subject order was the most frequent (Table 5.2). This may suggest that only coarse-grained generalizations are relevant in comparing corpus and experimental data, favoring a subject-object interpretation irrespective of the transitivity of the predicate.

However, the main clause data can be accounted for in a different way. The subject-object preference attested in the experiment may be due to the absence of transitivity information rather than to it being abstracted away from. In the main clause experiment, sentences were disambiguated right after the first NP by means of number information at the finite auxiliary (hadden / heeft in (1)). At this point in the sentence no information was provided concerning the number of NPs in the sentence. The clause could thus be continued with either an intransitive or (di)transitive predicate. At the point of disambiguation then, the only relevant grain-size was the coarser one, pooling intransitives and transitives. A frequency-based processing mechanism would therefore initially favor a subject-initial reading, even though it may be sensitive to more detailed properties of the predicate.

The subject-object preference attested for transitive main clauses in Experiment 1 therefore suggests either that only coarse-grained information plays a role, or that the level of abstraction depends on the information provided by the input: finer-grained transitivity information may be relevant if available.
3.2 The influence of the second NP: an apparent paradox

The experimental data on embedded clauses are more suitable to investigate the grain-size problem than the main clause data. In all but one of the experiments on embedded clauses, disambiguation took place at or after the second NP, rendering a transitive completion highly probable. The corpus data show that in all transitive welke-clauses, the object-subject reading is the most frequent. Comparing this to the experimental data on embedded clauses leads to an apparent paradox: different grain-sizes seem to apply depending on the type of second NP.

First consider the experimental data on welke-questions containing a pronoun as a second NP. When the pronoun was case-marked, a forced-choice completion and on-line judgment and reading tasks showed a preference for an object-subject order (Experiments 3-5). When the pronoun was not case-marked, an object-subject preference was eventually attested, as well (Experiment 6). This corresponds to the frequency data when counts are restricted to transitive welke-clauses, or even to transitive welke-clauses containing a pronoun: object-subject clauses are more frequent, especially when the second NP is a pronoun. This suggests that frequency-based parser must at least distinguish transitive from intransitive welke-questions in terms of frequency.

However, this level of abstraction does not seem to be the most appropriate for welke-questions containing a definite NP. The experimental data show that if any, a subject-object reading is preferred for these clauses. This is in contrast with the frequency data for transitive welke-questions: object-subject clauses are significantly more frequent, even if predicates containing a definite NP were considered separately. Hence, if there is a correspondence between frequency and processing, a coarser grain-size appears more appropriate.

The comparison between the experimental and the corpus data now leads to an apparent paradox. On the one hand, the data for welke-questions containing a pronoun suggest that transitive welke-questions are treated as a distinct clause type in terms of frequency. On the other hand, the comparison of the frequency and experimental data for welke-clauses containing a definite NP suggests that they are not.

3.3 Resolving the paradox

Experimental and corpus data for the welke-questions thus do not correspond in a straightforward way. This suggests that a grain-size according to which welke-questions are treated as a separate clause type is not appropriate. The relevant grain-size may either be coarser (more properties may be abstracted away from in classifying utterances), or finer (more details may be relevant) than the ones considered up to now. Alternatively, not one, but various levels of generalizations may play a role. Finally, the grain-size may have been appropriate, but frequency and comprehension may have failed to correspond for other reasons. I will discuss these issues in turn below.
3.3.1 A coarser grain-size

First, the frequency and comprehension data may correspond at a higher level of abstraction. In the present corpus study, only welke-questions were taken into consideration. However, such clauses may be non-distinct from other wh-questions, e.g. questions starting with wie (‘who’) or wat (‘what’). It may be that the relative frequency of subject-object versus object-subject wh-clauses in general corresponds exactly to the experimental data. This possibility cannot be tested at present, since no frequency studies on wie- and wat-clauses have been conducted yet.

Such an approach would not really solve the paradox, however. Note that on the one hand, a distinction must still be made between clauses containing a pronoun and ones containing a full definite NP as a second NP in order to account for the comprehension data. On the other hand, the properties of the first NP are abstracted away from: pronouns (wie) are considered equal to full NPs (welke-N). The paradox therefore remains: on the one hand, frequencies are stored separately depending on the type of NP involved; on the other hand, clauses are collapsed in terms of frequency irrespective of the type of NP.

3.3.2 A finer grain-size

An alternative solution is to refine the grain-size. Several subtypes of welke-questions may be distinguished on the basis of e.g. the animacy and number properties of the NPs. In addition, a wider context may be relevant. The relative frequencies of the subject-object and object-subject orders may differ depending on the givenness of the NP referents. Prince (1981, 1982) shows that given information is likely to be encoded as the subject of the clause (cf. Chapter 2). Hence, when the definite NP refers to given information, the object-subject order for welke-clauses may be most frequent; when the definite NP introduces new information into the discourse, a subject-object order may dominate.

If this is true, the discrepancy between the corpus and the comprehension data for the definite NP conditions may be accounted for in the following way. In the texts on which the corpus is based, the welke-question is part of a connected discourse. The definite NPs following the welke-phrase may have often referred to information that had already been introduced in the discourse. The object-subject order may therefore have been the most frequently attested in the corpus. In the experimental settings, on the other hand, the sentences were presented in isolation. The NPs thus always introduced new information. The subject-object order for these welke-clauses containing a full definite NP may then have been preferred on the basis of frequency. Whether frequency data indeed correspond to the experimental data when the discourse status of the NP referents is taken into consideration, and whether even more subclasses of welke-questions need to be distinguished remains to be investigated.
3.3.3 A mixed-grain approach

In the above I have discussed how the discrepancy between corpus and experimental data could be accounted for by assuming a coarser or finer grain-size. A different approach is to assume that not one, but several levels of generalization are relevant: abstract structural generalizations may be kept track of as well as frequency information tied to the specific properties of the NPs and the verbs. During on-line processing, these sources of information combine to yield a preference for one reading or the other (cf. MacDonald, Pearlmutter and Seidenberg, 1994).

In this view, the processing of a particular ambiguous welke-clause is not (or not only) influenced by the relative frequency of subject-object and object-subject welke-clauses in the language. Rather, various sources of information may play a role, such as the frequency of subject-object versus object-subject clauses in general; how often a welke-phrase is the subject or object of a clause; how often a pronoun, definite NP or other NP in general occurs as the subject or object of a clause, perhaps contingent upon the animacy properties and discourse status of the NP referent.

Some information may weigh more strongly than other. For instance, the bias towards having a pronoun occupy the subject position may be so strong that it overrides the general preference for a subject-object order. The subject bias for definite NPs may be somewhat weaker, and not strong enough to reverse the subject-object order favored by other sources of information.

A mixed-grain approach thus assumes that ambiguity resolution is determined by various factors, and not only on the basis of the relative frequency of the alternative readings of a particular structure. Experimental data concerning a particular structure therefore need not correspond to frequency data for that specific construction in a corpus. Of course, in order for such mixed-grain approaches to work, the relevant sources of information and the relative strength of their biases need to be specified and motivated. In Chapter 6, this approach will be further discussed in the section on constraint-based models.

3.3.4 Other explanations

The discrepancy between frequency and experimental data may also have been due to factors other than grain-size.

First, the size of the frequency difference between the subject-object and object-subject order may play a role. For the definite NP cases, the difference in frequency of the two alternatives may not have been large enough to have an effect on on-line processing. Table 5.5 shows that for the definite NPs only 65% of the clauses was object-subject. For the pronouns, which did show a object-subject preference in the reading times, this percentage was 92%. However, effects on reaction times have been found with even smaller differences between the alternative readings (cf. Cuetos, Mitchell and Corley, in press). Furthermore,
the small frequency advantage for the object-subject reading cannot account for the tendency for a subject-object reading in the definite NP conditions. The discrepancy between the corpus and the experimental data for the definite NP condition therefore cannot, or at least not fully, be attributed to the small difference in frequency between the two orders.

Second, the divergence between the experimental and the frequency data may be due to the nature of the corpus. The corpus used may not have been a representative sample of Dutch as it does not cover spoken language. It may be that in a corpus drawn from both spoken and written sources, welke-questions containing a definite NP occur more frequently as subject-object than as object-subject clauses, whereas the object-subject order is still the most frequent for clauses containing a pronoun. Whether this is true must await future research.

Third, the differences between frequency and comprehension data may be due to differences between perception and production strategies. The texts in the corpus may to a larger extent reflect production rather than comprehension processes. Production need not be driven by the same processes as comprehension (cf. e.g. Gibson, Schütze and Salomon, 1996). The most frequent structure in a corpus therefore need not correspond to the preferred structure attested in reading experiments: what is easy to produce need not always be easy to comprehend.

The frequency and experimental data may thus diverge for several reasons. The (apparent) discrepancy between frequency and reaction time data is not unique for Dutch welke-questions. See, for instance Hindle and Rooth (1993), Spivey-Knowlton and Sedivy (1995), Gibson, Schütze and Salomon (1996), and Schlesewsky, Fanselow, Kliegl and Krems (to appear) for other cases in which corpus and experimental data do not correspond in a straightforward way.

4 Summary

In this chapter the frequencies of occurrence of subject-object and object-subject welke-questions were investigated and compared with the experimental data. The frequency data confirm previous findings that subject-initial clauses are the most frequent in Dutch, at least, at an abstract level of generalization. Collapsed over all predicate types, a welke-phrase more frequently was the subject than the object of the clause. However, when only transitive welke-clauses were taken into consideration, the object-subject order was the most frequent. This corresponds to the experimental data (Experiment 1): main clause welke questions showed a robust preference for a subject-initial order when transitivity information was not available at the point of disambiguation.

The frequency data also provide additional support for the view that the nature of the second NP affects the order of subject and object. When the second NP is a pronoun, object-subject welke-clauses are far more frequent than subject-object questions. The frequency advantage for the object-subject order is significantly smaller when the second NP is a full definite NP.
Contrary to the prediction of at least some frequency-based parsing models, the frequency and experimental data did not correspond in a straightforward way, leading to an apparent paradox with respect to grain-size. Several solutions to this paradox were discussed. Most plausibly, more detailed (semantic, discourse) information has to be taken into consideration while tabulating frequencies. This is irrespective of whether a fixed grain-size or various levels of generalization are considered relevant. However, other explanations for the discrepancy between the corpus and the experimental data can be conceived of.