2

The linguistics of gender

This chapter explores grammatical gender as a linguistic phenomenon. First, I define gender in terms of agreement, and look at the parts of speech that can take gender agreement. Because it relates to assumptions underlying much psycholinguistic gender research, I also examine the reasons why gender systems are thought to emerge, change, and disappear. Then, I describe the gender system of Dutch. The frequent confusion about the number of genders in Dutch will be resolved by looking at the history of the system, and the role of pronominal reference therein. In addition, I report on three lexical-statistical analyses of the distribution of genders in the language. After having dealt with Dutch, I look at whether the genders of Dutch and other languages are more or less randomly assigned, or whether there is some system to it. In contrast to what many people think, regularities do indeed exist. Native speakers could in principle exploit such regularities to compute rather than memorize gender, at least in part. Although this should be taken into account as a possibility, I will also argue that it is by no means a necessary implication.

Grammatical gender

Throughout the preceding chapter, I have been relying on the reader's intuitions about (or professional knowledge of) grammatical gender. It is now time to be a little more specific about gender as a linguistic entity, and about the terminology that goes with it. Below, I will largely follow the comprehensive analysis of gender systems provided by Corbett (1991; see Corbett, 1994 for a summary). Much of the material will be exemplified only later, when I describe the Dutch gender system. For examples drawn from dozens of other languages, I refer to Corbett's original description.

Genders as agreement classes

Corbett begins his overview with a working definition of grammatical gender, which immediately sets it apart from natural gender:

> To understand what linguists mean by 'gender', a good starting point is Hockett's definition: 'Genders are classes of nouns reflected in the behavior of associated words' (1958: 231). A language may have two or more such classes of genders. The classification frequently corresponds to a real-world distinction of sex, at least in part, but often too it does not ('gender' derives etymologically from Latin *genus*, via Old French *gendre*, and originally meant 'kind' or 'sort'). The word 'gender' is not used for just a group of nouns but also for the whole category; thus we may say that a particular language has, say, three genders, masculine, feminine and neuter, and that the language has the category of gender. (Corbett, 1991, p. 1)

The fact that grammatical gender is a property of individual nouns, and not of the referents of those nouns, is expressed by the alternative terms 'lexical gender' and 'word gender'. The property shows up in the behavior of syntactically associated words. In German, for example, we can tell that the word 'Mädchen' falls in the set of neuter nouns because it takes the singular nominative definite article 'das'. In French, we can tell the gender of a word like 'pipe' from the fact that it goes with 'une' and 'grande', as in 'une grande pipe'. This behavior of associated words is usually called 'agreement', or 'concord':

> The term agreement commonly refers to some systematic covariance between a semantic or formal property of one element and a formal property of another. For example, adjectives may take some formal indication of the number and gender of the noun they modify. (Steele, 1978, p. 610; quoted in Corbett, 1991, p. 105)

It is not just that the grammatical gender of words like 'Mädchen' and 'pipe' shows up in agreement. Rather, as extensively argued by Corbett, agreement is also the determining criterion for grammatical gender:

> While nouns may be classified in various ways, only one type of classification counts as a gender system; it is one which is reflected beyond the nouns themselves in modification required of 'associated words'. (...) Saying that a language has three genders implies that there are three classes of nouns which can be distinguished syntactically by the agreements they take. (Corbett, 1991, p. 4)

The above examples all refer to languages with two or three genders, such as French or German. Traditionally, and because of the frequent (but partial) correlation with natural gender, the genders of Indo-European languages are usually referred to as masculine, feminine, and neuter. By defining gender in terms of agreement classes, however, it becomes clear that these are just

convenient labels that could in principle be replaced by 'class 1, 2 and 3'. Indeed, if a language has many more genders, as the Bantu languages mentioned before, this is how genders are often labelled.

Agreement targets

In the linguistic analysis of gender systems, the agreement patterns exhibited by associated words are used to infer the gender of a particular noun, as well as the total number of genders in the language. A noun like 'Mädchen' is said to be of neuter gender because of the particular 'agreement markers' found on associated words. Thus, in terms of the analysis, agreement patterns lead to noun gender. In terms of the generation of a surface structure, however, causality goes the other way: the gender of a noun is said to 'control' the form of particular 'agreement targets'. A familiar agreement target for gender is the definite article ('das Mädchen', 'la pipe', 'de ster'), but this is just one of many possibilities. Across more than 200 gender languages, Corbett (1991) has in fact observed a surprising variety of targets for gender agreement. Although usually not in a single language, a noun's gender can control the form of various attributive modifiers -- adjectives, demonstratives, definite and indefinite articles, numerals, and possessives -- as well as the form of verbs, predicative adjectives, relative pronouns, personal pronouns, adverbs, adpositions, and perhaps even complementizers. In terms of phrase structure location, targets for gender agreement need not be within the noun phrase headed by the 'controller noun', but can also be outside of it (NP-internal or -external agreement; cf. Lehmann, 1982).

Gender is a syntactic phenomenon because the agreement targets through which it shows up can only be defined in terms of their syntactic category (adjective, verb, etc.) and their syntactic relationship to the controller noun. It is however also a morphological phenomenon, because gender agreement is marked by inflectional devices. Whereas Indo-European languages usually mark gender on the agreement target by means of a suffix (e.g. French 'un petit chalet', 'une petite maison'), languages in other families may instead use prefixes, or a mixture of both. Although gender is typically marked by pre- or suffixes, it can also be realized by means of infixes, or by means of suppletion (as in the French articles 'le' and 'la', or the German nominative articles 'der', 'die', and 'das').[1] The total set

[1]I interpret an alternation such as 'le'/'la' as a true instance of (partial) suppletion, i.e. "the use of two or more distinct stems for forming the inflections of a single lexical item" (Trask, 1993, p. 270; see also Spencer, 1991, p. 128). Depending on one's theoretical framework, such alternations can also be said to involve two (or in the German case, three) separate lexical items.

of morphosyntactic forms used to mark a particular gender in some language is often called the 'paradigm' of that gender.[2] Because the morphosyntactic realization of gender frequently interacts with that of other linguistic categories such as number, case, person, or tense, a complete specification of the gender paradigms will often involve one or more of those categories as well (cf. German).

In some languages, the gender of a noun does not just show up in the morphology of its agreement targets, but also in the form of the noun itself. Such languages are said to have an 'overt' gender system. A particularly clear example of this comes from Swahili, a Bantu language: in 'ki-kapu ki-kubwa ki-moja ki-lianguka' ('7-basket 7-large 7-one 7-fell', 'one large basket fell'), the gender-marking prefix 'ki', which indicates that 'basket' belongs to gender class 7, also turns up on 'basket' itself (Welmers, 1973; quoted in Corbett, 1991, p. 117). In languages that have a 'covert' gender system, gender is -- by definition -- marked on the agreement targets, but it does not show up on the noun itself. A language can also be somewhere in between having a fully overt or covert gender system, with for example only a subset of the nouns being marked for their gender (e.g. Italian).

Distribution and diachrony of gender systems

Almost all of the gender examples given so far have been taken from Indo-European languages. Many languages in this family have gender. Some, like German, Icelandic, Serbo-Croat, and Russian, have three genders, traditionally labelled masculine, feminine, and neuter. Other Indo-European languages, such as Dutch, French, and Italian, have reduced the number to two in various ways. The Indo-European family may dominate linguistic analysis and its traditional gender systems terminology, but it is certainly not the only family with gender languages. Grammatical gender is also found in languages of, for example, the Caucasian, the Afro-Asiatic, the Niger-Kordofanian, the Dravidian, the Indo-Pacific, the Australian Aboriginal, and the Algonquian family. On the other hand, there is no gender in the Uralic family, several of the major families of Asia, and most of the languages in America.

[2]The term 'paradigm' can also be used to refer to the entire system of gender-marking forms, rather than just the subsystem for a single gender (Carstairs-McCarthy, 1994).

For many decades, linguists have asked themselves why, and how, gender systems have come into existence, why and how these systems change over time, and why, in the end, some languages lose their gender system again. Although the story is far from complete (see Corbett, 1991, pp. 310-318), it seems that gender systems arise from the use of nouns with classificatory possibilities, such as 'woman', 'man', and 'animal'. Such nouns may initially be used as classifiers, free forms that (often obligatorily) accompany some other noun in order to classify the latter's referent in terms of some important conceptual dimension. Consistent and repeated use of classifier forms (e.g. for 'woman') may gradually cause such forms to become attached to other parts of speech in the vicinity, such as adnominal adjectives. As a result of this, they will gradually be reanalyzed as morphologically marking some formal property of the noun whose referent was originally being classified (e.g. that it is a noun of 'feminine' gender). A variation on this theme involves the anaphoric use of classifiers, which may gradually lead to gender-marked demonstrative pronouns. The latter may in turn give rise to gender-marked articles and personal pronouns. And topicalized personal pronouns can in turn give rise to gender marking on verbs.

Once a gender system is in place, it is of course not immutable. For various reasons, a language is always on the move (De Saussure, 1916/1967; Crystal, 1987; McMahon, 1994), and there is no reason why its gender system would be an exception. As a language develops, its gender system can expand in terms of getting an extra gender, and the morphosyntactic markers upon which the system depends might be renewed or reorganized. A language can lose a gender, most often because the markers upon which it depends wear off for independent reasons, such as a gradual change in the language's sound system. The distribution of nouns across the genders in a language can change as well. A particular class may expand by taking in newly coined words, words borrowed from another language, or simply words that used to be in a 'rival' class. It may shrink because it loses words to another class, or because some of its words fall into disuse. Such changes of class size may seem fairly innocent, but they can conspire with other factors, notably the above-mentioned loss of gender markers (attrition), to result in the complete loss of a gender class. And if a language can lose a single gender, it can in the end lose its entire gender system.

Although the rise and decline of grammatical gender systems can thus be explained as a special concatenation of 'linguistic accidents', their journey does seem to follow a more general road that languages travel again and again. Beginning with the speakers' desire to be creative, yet limited by what the language has to offer, "forms which originally help build a coherent discourse become part of the syntax; grammaticalization then embeds them into the morphology, and subsequent phonological attrition fuses them into morphophonemic markers, then finally deletes them altogether." (McMahon, 1994, p. 168; after Givón, 1971; 1979; see also Bybee, 1994). It is not entirely

clear why language has this general tendency, but the fact that gender systems seem to "go with the flow" does suggest that their development is at least partly controlled by more general factors.

Why does gender exist?

Why do languages have grammatical gender at all? Although comprehensive in many other respects, Corbett (1991) is remarkably silent here. In view of the complexity of the issue and the scope of this thesis, I would rather stay away from it as well. Psycholinguistic research on the processing of gender in comprehension, however, is frequently taken to have implications for the raison d'être of gender (e.g. Grosjean Dommergues, Cornu, Guillelmon & Besson, 1994; Bates, Devescovi, Hernandez & Pizzamiglio, 1994), and to interpret such claims it is important to examine the assumptions behind them. A "large and frequently murky literature" (Greenberg, 1978) has been produced to explain the origin of gender (see I. Fodor, 1959, for a survey). For current purposes, however, it is enough to examine the available *types* of explanations, and the way they relate to potential functions of grammatical gender.

So, why does gender exist? Perhaps a constrained series of linguistic accidents, as for instance summarized by Corbett (1991, pp. 310-318), is all there is to it. Apparently, languages have a general tendency to take consistently used content and turn it into form. The reason why *gender* has travelled along this road, and has done so in several unrelated language families, might simply be that (1) people often want to talk about the fundamental categories in their world, such as male-female or animate-inanimate, but (2) they are constrained by their language to do this with certain forms only (e.g. the word for woman), and perhaps at certain structural positions only, so that (3) the resulting expressions become commonplace enough to *potentially* enter the morphosyntax of the language. Whether it does so or not will then depend on other accidents of linguistic history, accidents that will also determine the further development of a gender system once it has emerged. This kind of historical explanation would seem to be in line with modern views on language change (e.g. Aitchison, 1987; McMahon, 1994).

Unfortunately, though, it is not a particularly rewarding account. As historical linguists themselves admit, a post-hoc reconstruction of a series of linguistic accidents doesn't quite live up to the ideal of explanation as "strictly causal, universally valid covering laws which predict both *that* something will happen and *how*" (McMahon, 1994, p. 45). For many, it doesn't even meet the lower standard of providing "relief from puzzlement about some phenomenon" (Bach, 1974; quoted in Greenberg, 1979). After all, lots of things could have

come out of historical accidents. Instead, we only find a *limited* range of devices in the languages of the world, such as case, tense and, of course, gender.

Researchers have tried to explain this limited repertoire in two fundamentally different ways. In one approach, sometimes called **linguistic nativism**, it has been related to the biological make-up of the human species (e.g. Bickerton, 1984; Pinker & Bloom, 1990; Pinker, 1994). In terms of such a biological explanation, devices such as case and tense are part of our innate 'language instinct', which is why they show up in many unrelated languages. To my knowledge, however, this story has never actually been proposed for *gender*, at least not explicitly.[3] In order to explain the existence of gender in terms of an innate language instinct, one would have to show that it could have evolved in a Darwinian biological sense. Ultimately, then, one would have to argue that the presence of a grammatical gender device would contribute to the average reproductive success of individuals (Dawkins, 1986). Although such an argument may seem rather -- perhaps even very -- far-fetched, it has in fact been made for a whole range of syntactic devices (Pinker & Bloom, 1990; Pinker, 1994). If we wanted to apply it to gender, we would have to imagine that gender is such a useful device in verbal communication that it would have helped our ancestors in their struggle for survival and reproduction. That is, we would have to come up with some plausible communicative functions.

The other way to explain the existence of a limited range of linguistic devices, including gender, also rests on whether we can come up with communicative functions. Whereas theorists in the nativist tradition try to relate such functions to linguistic forms via the process of biological evolution, however, the theorists that subscribe to **linguistic functionalism** look for a more direct relationship. According to Bates and MacWhinney (1989), for example, it is not a language-specific biological instinct that makes languages all over the

[3]Because grammatical gender is typically absent from creole languages (Romaine, 1988), it has not made it to the 'Language Bioprogram' hypothesized by Bickerton (1984). However, two other leading nativists, Pinker and Bloom, do refer to gender as they list a large number of 'substantive language universals' that would reflect the human language instinct (Pinker & Bloom, 1990, pp. 713-714). Unfortunately, though, it is not clear whether they take gender *itself* to be one of the devices made available by this instinct, or whether they merely see it as one of the semantic distinctions that will naturally be encoded by *other* innate devices, notably pronouns and other anaphoric elements. In his recent book on the human language instinct, Pinker mentions a 16-level Bantu grammatical gender system to show that nonindustrialized people can have very sophisticated linguistic forms (Pinker 1994, p.27-28). This was meant to illustrate that the complexity of linguistic forms does not correlate with the level of cultural sophistication, but instead reflects a biological instinct shared by all cultures. As such, Pinker could be taken to suggest that gender is one of the devices made available by this instinct.

world converge on a similar repertoire, but the fact that there is simply a limited set of 'solutions' to the human communication problem:

> Human cognition and emotion provide the basic meanings and communicative intentions that any natural language must encode, together with a universal set of processing constraints that sharply delimit the way that meanings and intentions can be mapped onto a real-time stream of gestures and/or sounds. (Bates & MacWhinney, 1989, p. 6)

Researchers who subscribe to linguistic functionalism have actually tried to explain the existence of grammatical gender within that framework. To Bates and MacWhinney (1989), gender is a good example of a grammatical device that has (culturally) evolved to support the communicative process *itself*, rather than to express some communicative content. The particular function they have in mind here is referent tracking: "gender markers may be crucial in helping the listener to keep track of referents across a complex passage of discourse" (pp. 18-19). More recently, Bates et al. (1994) suggested that gender cues may also facilitate the recognition of words, a function that would help to explain "why so many of the world's languages persist in the use of a costly linguistic device that serves no obvious communicative function" (p. 3).

In contrast to the historical explanation that I discussed before ("linguistic accidentalism"), both alternative types of explanation, linguistic nativism and linguistic functionalism, relate the existence of gender to its hypothesized communicative functions. What functions have been proposed in the literature? As already mentioned, researchers have often pointed out that grammatical gender can disambiguate anaphoric or deictic referential constructions, and can as such help to keep track of the referents in a discourse (e.g. Zubin & Köpcke, 1981; Köpcke & Zubin, 1984; Mills, 1986; Bates & MacWhinney, 1989; Corbett, 1991). Somewhat less often, it has been suggested that gender can also help to process other types of constructions, such as nested noun phrases or compound nouns (e.g. Köpcke & Zubin, 1984; Mills, 1986; Wijnen & Deutsch, 1987). A related suggestion is that, by merely showing which words go together, gender increases the syntagmatic cohesion of a sentence, which may in turn facilitate its processing (Desrochers, 1986). Recently, psycholinguists have begun to focus on the potential contribution of gender to the word recognition process (e.g. Grosjean et al., 1994; Colé & Segui, 1994; Bates et al., 1994). Without committing themselves to a particular processing locus, others have suggested that gender might help to set up 'expectations' about what the speaker is going to say next (Köpcke & Zubin, 1984; Mills, 1986). In addition to all these rather prosaic functions, Corbett (1991) has pointed out that gender is sometimes used to mark status, to show respect or a lack of it, and to display affection. A final function, one that seems to have been overlooked so far, is related to the fact that the very *redundancy* of gender systems actually provides the listener or reader

with a means for error detection (see Miller, 1991, pp. 24-25, for this interpretation of redundancy in language). In principle, the gender on a determiner doesn't tell us anything that we could not find out from the noun: the correctly gender-marked NP 'het volk', 'the people', is not more informative than the ficticious non-marked 'det volk'. However, whereas it is very difficult to assess whether something has gone wrong in the transmission of a numeric string like '45 5085915 6297 05541...', a Dutch native speaker can easily see that something has gone wrong with a linguistic string perceived as 'de Engelse volk heeft...' (when this particular agreement violation prompted me to reanalyze the input string, the actual word turned out to be the de-word '*Vogue*', a well-known fashion journal).

It is not unlikely that researchers will come up with other possible functions of grammatical gender. Some of the above proposals, such as the capacity of gender to disambiguate syntactic constructions, are obviously correct (although perhaps not for every gender language); if the final interpretation of a sentence hinges on grammatical gender, the latter simply must have an impact in sentence processing somewhere. Other proposals, such as that gender might help in word recognition, remain to be verified (see Chapter 3). For a correct interpretation of such research, however, it is important to keep in mind that an established function of gender, interesting as it may be, need not have *anything* to do with the reasons why a gender system has emerged or persists in the language at hand. If grammatical gender turned out to facilitate the recognition of words, for example, it might be that this is simply a *side effect* of the way the lexical nature of gender interacts with general operating principles of the word recognition process; the gender system itself may have arisen, and may persist, for completely different reasons. Maybe it is there because of some other function (exerting its influence along nativist or functionalist lines). Or, to return to the beginning of this section, maybe it is just there because of a series of linguistic accidents, controlled by general principles of language change. We cannot exclude the possibility that, once they have sprung into existence, gender systems are simply tolerated because they do no harm (see also Carstairs-McCarthy, 1994).

One important aspect of grammatical gender systems has not yet been addressed. Many people who have given more than a moment's thought to such systems are convinced that, although agreement is systematically related to the gender of a noun, this gender itself is often arbitrary, i.e., not systematically related to other properties of the noun or its referent. Why is the German word for spoon masculine, for fork feminine, and for knife neuter? And why is the word for house masculine in Russian, feminine in French, and neuter in Tamil? Both within and across languages, the assignment of nouns to genders does indeed appear to be a largely random affair. However, Corbett (1991) and a number of

other linguists have recently argued for just the opposite. If they are correct in their claim that the gender of a noun can very often be derived from other properties of the noun or its referent, this has some interesting implications for the way gender may be processed by the language user. I will turn to this important issue in the last section of this chapter. First, however, I want to have a look at the grammatical gender system of Dutch.

The Dutch grammatical gender system

Those with metalinguistic knowledge of Dutch may have been surprised to see the language described as having *two* genders, associated with de-words and het-words respectively. Didn't Dutch have three genders? Take any Dutch dictionary, and the nouns therein will be marked with **m.**, **v.**, and **o.**, Dutch abbreviations of 'mannelijk' (masculine), 'vrouwelijk' (feminine), and 'onzijdig' (neuter). Following a large number of -- mainly Dutch -- publications on this matter, I will argue below that, the masculine-feminine gender distinction is an artificial partitioning, still being enforced by dictionaries and normative grammars, but no longer alive in the spontaneous use of northern Standard Dutch, the Standard Dutch as spoken in most of the Netherlands.[4]

Genders

The nouns of colloquial northern Standard Dutch are distributed across two grammatical genders (van Haeringen, 1954; Geerts, 1968; Dekeyser, 1980; Geerts, Haeseryn, de Rooij & van den Toorn, 1984; Donaldson, 1987; Geerts, 1988; van Beurden & Nijen-Twilhaar, 1990; Verhoeven, 1990; Kooij, 1992, van den Toorn, 1992; van der Wal, 1992). Nouns that take the singular definite article 'het', such as 'het huis', are called **het-words**, and can be referred to as having the 'het-gender' or the 'neuter' gender. Nouns that take the singular definite article

[4]In the remainder, northern Standard Dutch will usually be abbreviated to 'Dutch'. I will not be concerned with the variety of Standard Dutch spoken in Belgium and the most southern areas of the Netherlands (southern Standard Dutch), nor with dialects spoken in either country. From the perspective of gender systems, these are important restrictions (e.g. Dekeyser, 1980). The grammar of northern Standard Dutch has been made very accessible by Donaldson (1987). For a comprehensive grammar of Standard Dutch, I refer to Geerts, Haseryn, de Rooij & van den Toorn (1984; 1300 pages, and in Dutch). A more gentle 20-page English introduction to Standard Dutch has been provided by Kooij (1987; see also Kooij 1992).

'de', such as 'de ster', are called **de-words**, and can be referred to as being of the 'de-gender', 'common' gender or 'non-neuter' gender.[5]

From the perspective of psycholinguistic research, one of the most important things to know about these two gender classes is their relative size. How many Dutch nouns are de-words, and how many are het-words? Before the advent of computerized lexical databases, it was very hard to get a reliable estimate (see Czochralski, 1983, for an example of the use of very anecdotal evidence). An early estimate by Tuinman (1967; quoted in Geerts, 1975) was that roughly 75% of Dutch nouns were de-words, and only 25% were het-words. A similar ratio has been reported by Deutsch and Wijnen (1985). And, using the computerized Woordenlijst van de Nederlandse Taal of 1954 as his source, Frieke (1988) observed that no less than 81% of 2943 monosyllabic nouns were de-words, against only 19% het-words.

Going by these estimates, it seems that Dutch has more de-words than het-words, possibly even three to four times as many. However, now that we have a large computerized lexical database of Dutch, the CELEX Dutch lexicon V3.1, with 130,788 word entries based on a running text corpus of some 42 million words (the INL corpus; Burnage, 1990), we can be much more precise. To that effect, I have carried out a number of lexical-statistical analyses. Because they are fairly detailed, and would as such disrupt my sketch of the Dutch gender system, I will describe them in a later section. They will confirm that Dutch has more de-words than het-words, but will also show that the overall ratio varies from 3:1 to 2:1, depending on how you count your words.

The vast majority of Dutch nouns is *either* a de-word *or* a het-word. A small number of so-called 'double-gender nouns', however, take both de-word and het-word agreements. Thus, it is possible to say 'de brok' or 'het brok' (both 'the lump'), 'de schort' or 'het schort' ('the apron'), 'de aanrecht' or 'het aanrecht' ('the sink'), and 'de matras' or 'het matras' ('the mattress'). Often, one of the two genders is preferred (Geerts et al., 1984). The essence of such double-gender nouns, however, is that the alternative gender is always allowed as well, and that it does not change the semantics. Many double-gender nouns may in fact be in the process of changing from one gender to the other (de Rooij, 1977).

Double-gender nouns should not be confused with what I will call 'different-gender homonyms'. Take the word-form 'jacht', for example. Although it allows for both genders, the semantics reveal that there are simply two different nouns involved: 'de jacht' means 'the hunting', whereas 'het jacht' means 'the yacht'. Thus, 'jacht' is not a double-gender noun, but a different-gender

[5]These somewhat awkward gender terms for de-words derive from the fact that, as will be seen below, this class is a merger of the historical masculine and feminine genders.

homonym, with *two* nouns that happen to share the same word-form but not the same gender.

Agreement targets

One of the targets for Dutch gender agreement has already been used in several examples above: the **definite article**. As shown in Table 2.1, there are a number of other agreement targets. In their attributive use, the proximal and distal **demonstrative pronouns**, the **possessive pronoun**, the **interrogative pronoun**, and a number of **indefinite pronouns** all show gender agreement. Of the independently used pronouns, only the **relative pronoun** is reliably marked for gender (although the use of 'die' is perhaps no longer exclusively limited to de-word antecedents, especially if the antecedent is relatively distant and/or has a human referent; cf. Verhoeven, 1990). In indefinite noun phrases, many attributive **adjectives** also inflect for gender. The indefinite adjective inflection, if present, always marks common gender with an '-e' suffix, and neuter gender with a zero suffix.[6]

Taken together, the forms of the de-word paradigm are sometimes referred to as [+e] forms, and those of the het-word paradigm as [-e] forms (e.g. Fletcher, 1987).[7]

---

[6]The rules for adjective inflection are actually rather complex. Apart from indefinite NPs that have no article at all ('klein huis') or that begin with an indefinite article ('een klein huis'), the NPs that begin with 'geen', 'veel', 'weinig', 'wat', 'zo'n', 'zulk', 'ieder', 'enig', 'menig', 'welk', 'wat voor', 'genoeg', or 'allerlei' may also have a gender-inflected adjective (Fontein & Pescher - ter Meer, 1985). Whether, in all such NPs, an adjective actually does mark gender will also depend on a number of other factors that are beyond the scope of this thesis (but see Donaldson, 1987, pp. 72-75; Geerts et al., 1984, pp. 322-331). In a *definite* noun phrase, however, the adjective *never* marks gender, as in 'de kleine ster' and 'het kleine huis', or 'die kleine ster' and 'dat kleine huis'.

[7]The gender system of present-day northern Standard Dutch uses only a subset of the possible agreement targets observed by Corbett (1991). The particular selection made by Dutch is perhaps best appreciated if we try to imagine agreement targets that have *not* been used. Thus, like several other gender languages, Dutch might have had indefinite article agreement, e.g. *'een**e** ster', 'een huis' ('an X'), numeral agreement, e.g. *'vier**e** sterren', 'vier huizen' ('four Xs'), verb agreement, e.g. *'de ster staat**e** daar', 'het huis staat hier' ('the X stands here'), predicative adjective agreement, e.g. *'de ster is klein**e**', 'het huis is klein' ('the X is small'), subject-object agreement on a personal pronoun, e.g. *'ik**ke** zie een ster', 'ik zie een huis' ('I see an X'), preposition agreement, e.g. *'achter**e** de ster', 'achter het huis' ('behind the X'), and complementizer agreement, e.g. *'dat**te** de ster weg is', 'dat het huis weg is' ('that the X is gone').

	de-words *common gender*	***het-words*** *neuter gender*	*English equivalent*
definite article	de ster	het huis	the star/house
demonstrative pronouns	deze ster die ster	dit huis dat huis	this star/house that star/house
possessive pronoun	onze ster	ons huis	our star/house
interrogative pronoun	welke ster?	welk huis?	which star/house?
indefinite pronouns	elke ster iedere ster menige ster	elk huis ieder huis menig huis	each star/house every star/house many a star/house
relative pronoun	de ster die ...	het huis dat ... het huis wat ...	the star/house that ...
adjectives in indefinite NPs	(een) kleine ster (een) rode ster ...	(een) klein huis (een) rood huis ...	(a) small star/house (a) red star/house ...
but...	de rode ster	het rode huis	the red star/house

Table 2.1 Agreement targets in the Dutch gender system.

The gender contrasts in Table 2.1 are made for singular nouns only. In the plural, gender is never marked. The inflectional paradigm for plural de-words *and* het-words is actually identical to that for singular de-words. As a result, there is considerable syncretism within the system. The definite article 'de', for example, can be the determiner of a singular de-word, of a plural de-word, and of a plural het-word. For a good understanding of how the gender system is embedded in the Dutch language, it is also important to know that several forms in the gender paradigm also realize some completely different part of speech. The most important ones are 'het', also the form of an expletive pronoun ('het regent', 'it rains') and of a personal pronoun ('Ik zag het zonet nog', 'I saw it just now'), and 'dat', which is also the main Dutch complementizer ('Ik weet dat zij weg is', 'I know (that) she is gone').

Diachrony and the pronominal system

Like English and German, Standard Dutch belongs to the West Germanic branch of the Indo-European language family. Proto-Indo-European, the reconstructed parent language of this family, is believed to have been spoken before 3000 BC (Crystal, 1987), and also held to be the source of the masculine-feminine-neuter distinction found in many of its offspring (Corbett, 1991). Modern German has clearly preserved a version of this three-gender system. On the other hand, whereas Old English also had a three-gender system, Modern English has practically -- if not completely -- lost it (Dekeyser, 1980).

With respect to its gender system, and its inflectional morphology in general, the Dutch language is often described as developing from a state that resembles Modern German to one that resembles Modern English (Van Haeringen, 1954; Dekeyser, 1980; Kooij, 1987; Geerts, 1988; Vandeputte, Vincent & Hermans, 1991). Indeed, the gender system of Middle Dutch (1100-1500) still looked very much like the Modern German one, having three genders crossed with four cases, and gender being marked on a wide variety of agreement targets (including indefinite articles; see Burridge, 1993, Appendix 1). Even at this time, however, the system was already showing signs of collapse. Geerts (1988) suggests that the transition towards a two-gender system may in fact have started at the end of the period of Old Dutch, when full vowels in word-final syllables were reducing to schwas, and the adnominal inflectional system thereby began to lose its markers.

It is generally thought that northern Standard Dutch is now at the end of this transition, and that the masculine and feminine gender classes have merged into a single 'common' or 'non-neuter' gender (Simons, 1920; Verdenius, 1946; van Haeringen, 1954; Geerts, 1968; Dekeyser, 1980; Nienhuys, 1983; Fontein & Pescher - ter Meer, 1985; Crystal, 1987; Donaldson, 1987; Geerts, 1988; van Beurden & Nijen-Twilhaar; 1990; Verhoeven, 1990; Kooij, 1992; van den Toorn, 1992; van der Wal, 1992). Admittedly, as with many issues of language change and language variation, it is very difficult to establish whether the three genders have disappeared from the colloquial speech of *all* speakers of northern Standard Dutch, and also where exactly northern Standard Dutch begins and southern Standard Dutch ends. We can safely assume, however, that most of the native speakers of Standard Dutch in the Netherlands have only de-word and het-word agreement in their spontaneous speech. And, although it is often said that the masculine-feminine distinction is alive and well in southern Standard Dutch (e.g. WNT, 1954; Donaldson, 1987), a number of linguists have in fact argued that the distinction has begun to collapse there as well (Geerts, 1968; Dekeyser, 1980; Geerts, 1988; Verhoeven, 1990).

Readers with some metalinguistic knowledge of Dutch may wonder why I have until now almost completely ignored the language's pronominal reference system. Isn't the masculine-feminine distinction still being expressed in third person singular pronouns, such as the personal pronoun forms 'hij' ('he'), zij ('she'), and 'het' ('it'), or the possessive pronoun forms 'zijn' ('his') and 'haar' ('her')? Why else would dictionaries specify the gender of every noun as **m.**, **v.**, or **o.**, i.e. as masculine, feminine, or neuter? And why else would native speakers of Dutch, as they write, bother to use such dictionaries to see whether a word is masculine or feminine?

The simple truth is that they bother because they have been told to bother, by prescriptive grammarians, and even by the Dutch government! Ever since the 17th century, Dutch grammarians have tried to counteract the gradual collapse of the masculine-feminine distinction (Geerts, 1988; see also van der Wal, 1992, or de Vries, Willemyns & Burger, 1994). With the inflectional morphology of classical languages as the ideal, the evolution towards a simple two-gender system was seen as degradation and decay, and believed to result from the fact that certain northern Dutch language users didn't know better or didn't care enough. In order to stop this undesirable development, grammarians frequently reminded the people of the 'correct' gender system by distributing lists of Dutch words together with their (masculine, feminine, or neuter) gender. At first, these were relatively small-scale private initiatives. In 1700, for example, David Hoogstraten published *Aenmerkingen over de geslachten der zelfstandige naemwoorden*, which was simply a compilation of the nouns and their genders as used by the two prominent Dutch writers Hooft and Vondel (de Vries, Willemyns & Burger, 1994). Over time, however, the concern for language standardization increased, and it became an issue for the national government. Amongst other things, the fact that southern Dutch did not (yet) evolve towards a two-gender system was felt to threaten the unity of the Dutch language. In the 19th century, therefore, grammatical gender was included in an official spelling revision, and as such laid down in the *Woordenlijst der Nederlandsche Taal* (de Vries & te Winkel, 1866). A revised version, the *Woordenlijst van de Nederlandse Taal*, more commonly known as het *Groene Boekje*, first appeared in 1954, and was reprinted without modifications in 1984 (WNT, 1954). Apart from a now very long list of words with their genders, it also contained the rules for 'correct' pronominal reference. Although it allowed for some variability in the use of 'masculine' and 'feminine' anaphoric pronouns, people were by and large supposed to follow the WNT of 1954. In fact, it was backed up by a Belgian and a Dutch spelling law.[8]

[8]The spirit of this language planning exercise is probably illustrated best by the way its goal was stated in 1954 (and 1984!): "A list of words is not there to put good stylists and linguistic artists under restraint, but rather to supply the 'common' people, as the Report-1936

Of course, genders that have been lost from the mental lexicon of native speakers will simply not come back by decree. And the related system of pronominal reference, no longer part of native speaker competence, is not easily reinstalled either. In the light of modern ideas about language change (e.g. McMahon, 1994; Pinker, 1994), one would indeed not expect prescriptive laws to be very effective. This particular instance of language planning, however, turned out to have a rather nasty side effect. Although native speakers of Dutch were, by and large, no longer able to use three genders in their spontaneous speech, many of them could be persuaded to at least make the effort *in their writing* (which was considered to be more important anyway). This has led to the unfortunate situation that, whereas it has vanished from spontaneous speech, the historical three-gender system has been *artificially* preserved in written northern Standard Dutch (Simons, 1920; Verdenius, 1946; van Haeringen, 1954; Geerts, 1968; 1988, Verhoeven, 1990; van Sterkenburg, 1991; van der Wal, 1992). As far as gender is concerned, the written variety of Dutch has even been described as a 'foreign' language (e.g. van Haeringen, 1954).[9]

Although it is clear that native speakers of northern Standard Dutch are not following the WNT rules in their *spontaneous* use of third person pronouns like 'hij', 'zij', and 'het', it is not entirely clear what it is they do instead. Van Haeringen (1954) has argued that pronominal reference is controlled by two orthogonal factors: the grammatical gender of the words involved (common or neuter), and the natural gender of the referents (male, female, or none at all). Whereas natural gender would be the dominant controller of personal and possesive pronoun forms, grammatical gender would be the dominant controller of relative and, to a lesser extent, independently used demonstrative pronouns (and in full control of adnominal forms, of course). In the framework of

---

put it, with a norm. By 'common people' we mean tradesmen and small civil servants, whose independent sense of style and language we shouldn't trust too much. We are also thinking of the difficult task of schoolmasters. (...) Young children in general and dialect-speaking children in particular cannot be assumed to have the linguistic feeling and the firm command of language needed to find their way in the precarious 'pronominal problems' of Dutch without guidance or prescription. All these simple folks need rules." (translated from WNT, 1954; 1984; p. xxii). In the revised version of the WNT, the *Herziene Woordenlijst van de Nederlandse Taal* (HWNT, 1990), this passage has disappeared.

[9]Whereas it still propagates the historical three-gender system, the most recent revision of the WNT (HWNT, 1990, p. 40) appears to have given up on the spoken variety of Dutch. It now explicitly aims at written Dutch, and acknowledges that things may be different in the spoken language. In the spirit of the 90s, though, it also explicitly allows people to disobey the rules, even in their writing. Hopefully, the next edition will give up on the three-gender system altogether.

Government and Binding theory, however, Verhoeven (1990) has recently made a more radical proposal: with the possible exception of the relative pronoun, *none* of the independently used pronouns would be under control of the grammatical gender system (not even the two-gender system). Pronominal reference would instead be determined by semantic factors, not only including the natural sex of the referent (e.g., 'Het meisje heeft haar beleid veranderd', 'The girl has changed her policy'), but for example also whether it is a collective entity or not (e.g., 'Het kabinet heeft haar beleid veranderd', 'The cabinet has changed its policy'). Further linguistic and psycholinguistic research is obviously needed to evaluate these proposals. But, whereas the precise influence of the two-gender system in pronominal reference is now subject to debate, linguists do agree that there is no *three*-gender system at work anywhere in present-day spontaneously spoken northern Standard Dutch.

Before leaving the topic of pronominal reference, I would like to point out that, even though the choice of a Dutch personal or possessive 3rd person singular pronoun is, for a human referent, controlled by *natural* rather than grammatical gender, this natural gender does have *grammatical* implications (van Haeringen, 1954). Just like English, Dutch has grammaticized a natural gender distinction in some of its pronouns, such that a native speaker who wishes to use such a pronoun *must* specify the natural gender of its human referent -- even if he or she (!) would rather leave it unspecified. At this point, confusion may easily arise, because 'grammaticized natural gender' does sound very much like 'grammatical gender'. Under a slightly different construal of the term, 'grammatical gender' could indeed be taken to cover such obligatory pronominal distinctions. In this thesis, however, it is defined as a matter of agreement with a formal property of nouns, and not with a biological property of referents. The fact that the latter property may *also* have become grammaticized in some parts of the language, as in the Dutch 'hij'/'zij'/'het' distinctions, does not by itself make such distinctions part of a grammatical gender system. They may be, but they need not be: the criterion is whether these pronouns agree with the gender of a (possibly implicit) antecedent *noun*, or whether they go their own way (see also Trask, 1993; p. 115).

Relative distribution of the two genders

With the masculine-feminine distinction out of the way, I can safely resort to counting de-words and het-words. Early estimates discussed before suggested that Dutch has at least three times as many de-words as het-words. But, in spite of their convergence, they leave much to be desired. The word samples upon which they were based are all rather small, and, in the case of Frieke's (1988)

analysis, limited to monosyllabic words.[10] More importantly, these estimates all involve *type* counts. That is, they estimate the number of de- and het-words that would be listed in a complete dictionary of Dutch (which of course will never exist). Although this is almost certainly a relevant feature of the 'gender environment' of a native speaker of Dutch, there are other, equally relevant ways to count de- and het-words. For example, how often does a native speaker encounter individual *tokens* of de- and het-words in spoken or written discourse? That is, given any one noun in a piece of running text, what is the probability that it will be a de-word rather than a het-word? By itself, the probability of encountering a de-word rather than a het-word in the dictionary does not say anything about the probability of encountering a de-word in running text, because the latter also depends on how often every de- and het-word is actually being used in the language.

The CELEX computerized lexical database of Dutch (version 3.1, Burnage, 1990) allows us to improve upon earlier estimates, both in terms of sample size and in terms of doing an additional token analysis. With its 130,788 word entries based on a 42 million word corpus of sampled texts, the lexicon itself can in fact hardly be called a sample. Of course, as a snapshot of written Dutch in the 1980s, it will inevitably miss out on many newly created or borrowed words, and on words that are limited to spoken language use. Still, for a reliable estimate of the relative proportion of de- and het-words in the language, it appears to be more than sufficient. Moreover, because every word entry in the database has been annotated with a corpus-based token frequency, it allows for both a 'dictionary-based' type count and a (written) 'text-based' token count. We can therefore look at both aspects of the gender environment of native speakers of Dutch.

In fact, CELEX offers much more than just 'dictionary entries' and their token frequencies. In Dutch, a noun such as 'huis' can have four inflectional variants: the singular 'huis', the plural 'huizen', the singular diminutive 'huisje', and the plural diminutive 'huisjes'.[11] In the CELEX Dutch **morphosyntactic word** lexicon, all four variants are explicitly represented, and each of these types has its own token frequency. In the CELEX **lemma** lexicon, there is just the word entry 'huis', and the token frequency of this lemma type is simply the sum of the token frequencies of each of its morphosyntactic word types. In the first two analyses below, I will make use of the CELEX lemma lexicon. In the third,

[10]To be fair, I should mention here that it was not Frieke's goal to estimate the relative distribution of de- and het-words.

[11]Although the diminutive is treated as an inflectional variant in CELEX, there is a linguistic argument for it to be a derivation, i.e., a different word. I will return to this below.

however, I will exploit the extra information encoded in the CELEX morphosyntactic word lexicon.

Lemma types

In order to verify the earlier estimates of the relative distribution of de- and het-words, all of which most likely involved dictionary entries, I will begin with a lemma type count in the CELEX Dutch lemma lexicon. The results of this analysis are displayed in Table 2.2. After excluding the rather arbitrary set of proper nouns in this database, I counted 92,628 common noun entries.[12] Of this set, 72% were unambiguously classified as de-words, and 27% as het-words, a de-het ratio of almost 3:1. Thus, although the lemma type distribution is slightly less skewed, the earlier estimates were close: approximately every fourth noun in the 'almost complete dictionary' of Dutch is a het-word, and the other three are de-words.

CELEX also allows us to look at the distribution within the subsets of monomorphemic and morphologically complex nouns. Dutch derivational morphology appears to have a large influence on the gender of a complex noun. In nominal compounds, for example, the gender of the rightmost part determines the gender of the compound: in 'de veldsport', 'the outdoor sport', it is the de-word 'sport' that determines the resulting gender, whereas in 'het sportveld', 'the sports area', it is the het-word 'veld' (Trommelen & Zonneveld, 1986). For nouns formed by suffixation, it has likewise been argued that the resulting gender is usually determined by the rightmost suffix (Trommelen & Zonneveld, 1986). This argument is supported by many suffix-based gender regularities, such as that words formed with the suffix '-heid' (e.g. 'vrijheid', 'freedom') are all de-words. Taken together, compounding and suffixation could cause the de-het ratio of morphologically complex words to substantially diverge from that of monomorphemic words.

[12]Within the CELEX database, alternative spellings and alternative morphological parses of a word are actually listed as separate lemmas. From the perspective of a de-het count, these alternatives are unwarranted duplications. They have therefore been excluded from all below analyses.

The set of common noun lemmas contains 72,592 (78%) morphologically complex nouns, and the ratio there is again almost 3:1. Within the much smaller subset of 6,349 monomorphemic words (7%), however, the de-het ratio is almost 4:1. The latter ratio confirms the monomorphemic estimate mentioned before (Frieke, 1988). In addition to morphologically simple and complex words, CELEX distinguishes three other morphological subcategories: lexicalized flections, morphologically irrelevant words, and morphologically unanalyzed words (Burnage, 1990, p. 3.56-3.59). Lexicalized flections are inflectional variants that have taken on their own meaning, such as 'avondje' (literally 'small evening', but usually a 'social evening'). Morphologically irrelevant words have a stem that does not allow for morphological analysis, e.g. because it involves a proper noun ('leninisme', 'leninism'). Morphologically unanalyzed words defy satisfactory analysis for a variety of other reasons, e.g. because of a classical affix ('genus', 'gender'). To avoid unnecessary detail in my lexical statistics, I have collapsed all this into a single 'other' category.

	de	*het*	*rest*	*total*	*de-het ratio*
monomorphemic	4,982	1,290	77	6,349	3.9
	78.5%	20.3%	1.2%	(6.9%)	
complex	52,506	19,372	714	72,592	2.7
	72.3%	26.7%	0.9%	(78.4%)	
other	9,120	4,155	412	13,687	2.2
	66.6%	30.4%	3.0%	(14.8%)	
total	66,608	24,817	1,203	92,628	2.7
	71.9%	26.8%	1.3%		

Table 2.2 The distribution of gender over all *common noun types in the CELEX lemma lexicon*. Shown are the number of de-word types (*de*), het-word types (*het*), and unclassifiable noun types (*rest*), each also as a percentage of the total number of types (*total*), as well as the number of de-word types divided by the number of het-word types (*de-het ratio*), within the monomorphemic noun stratum, the morphologically complex noun stratum, the remaining noun stratum, and the total noun stock (bracketed percentages express the relative size of a stratum within the total noun stock). Inaccuracies in marginal percentages are the result of rounding.

Lemma tokens

The above lemma type count has confirmed earlier estimates: about every fourth noun in the Dutch lemma lexicon, a dictionary with almost 100,000 nouns, turned out to be a het-word. However, as already mentioned, this doesn't imply that every fourth noun in running text is going to be a het-word. To establish this, we need to count all the de- and het-word tokens in a representative sample of Dutch text. Fortunately, most of the work has already been done. In the CELEX lemma lexicon, every lemma has been annotated with its frequency of occurrence (across all inflectional variants) in the INL corpus, a sample of 42,380,000 words of running text taken from 835 different written sources (Burnage, 1990). Thus, the only thing left to do is add up the frequency counts of all de-word lemmas, and compare the result to the summed frequency counts of the het-words.

	de	***het***	***rest***	***total***	***de-het ratio***
monomorphemic	2,464 k	1,037 k	6 k	3,507 k	2.4
	70.3%	29.6%	0.2%	(47.1%)	
complex	1,710 k	500 k	6 k	2,216 k	3.4
	77.2%	22.6%	0.3%	(29.7%)	
other	896 k	813 k	19 k	1,727 k	1.1
	51.9%	47.1%	1.1%	(23.2%)	
total	5,070 k	2,350 k	31 k	7,450 k	2.2
	68.0%	31.5%	0.4%		

Table 2.3 The distribution of gender over all *common noun tokens in the INL text corpus*. Shown are the number of de-word tokens (*de*), het-word tokens (*het*), and unclassifiable noun tokens (*rest*), each also as a percentage of the total number of tokens (*total*), as well as the number of de-word tokens divided by the number of het-word tokens (*de-het ratio*), within the monomorphemic noun stratum, the morphologically complex noun stratum, the remaining noun stratum, and the total noun stock (bracketed percentages express the relative size of a stratum within the total noun stock). All absolute token counts are in thousands (k = 1,000), and have been computed from the CELEX lemma frequency counts. Inaccuracies in marginal percentages are the result of rounding.

The results of this token count, for which I used the same set of 92,628 common nouns as before, are displayed in Table 2.3. Of the 7,450,089 noun tokens involved (18% of the total INL corpus), 68% were de-word tokens, against 32% het-word tokens. This is a de-het ratio of slightly more than 2:1. That is, of every noun encountered in running text, about every *third* turned out to be a het-word. Clearly, het-words have a higher average token frequency, which reduces the asymmetry seen in the earlier dictionary counts.

Singular noun tokens with diminutive correction

In the above analyses, I have explored two undoubtedly relevant aspects of the written 'gender environment' of Dutch native speakers: how many of the words encountered in a dictionary, and how many of the words encountered in running text will be of common and neuter gender? In terms of psycholinguistic implications, the 3:1 dictionary count could be taken to estimate the proportion of de- and het-words in the average native speaker's mental lexicon (although, depending on the way morphologically complex words are represented, one might want to argue that the 4:1 monomorphemic words ratio is the more appropriate estimate). The 2:1 running text count, on the other hand, may well be an environmental asymmetry that native speakers unconsciously pick up on, and that might lead to biases in their processing of gender (in real life or in gender-related laboratory tasks).

We cannot simply assume, however, that the above running text count is a good indication of the relative salience of the two genders in Dutch language use. The main reason is that, although a particular word token may be marked as a de-word or a het-word in the dictionary, its gender need not be obvious from the surrounding corpus text. For example, the above lemma token counts also include the plural occurrences of a noun. In plural contexts, gender is never marked. In terms of the (implicit) salience of the two genders of Dutch, it is therefore perhaps better to look at *singular* noun tokens only. These are the ones that are likely to be marked for gender in the surrounding text.

Related to the above, there is something else that I have ignored so far. In the CELEX database, a diminutive like 'huisje', 'small house', is taken as an inflectional variant of the lemma 'huis'. This means that all the occurrences of 'huisje', e.g in the phrase 'het kleine huisje', contribute to the token frequency of 'huis'. In this case, the diminutive takes neuter agreement, just like the noun lemma it belongs to. In Dutch, however, *all* diminutives formed with the suffix '-je' and its allomorphs '-kje', '-pje', '-tje' and '-etje' take neuter agreement. Thus, the diminutive 'sterretje', 'small star', takes neuter agreement even though 'ster' itself is of common gender. And because CELEX treats 'sterretje' as an inflectional variant of 'ster', noun phrases such as 'het kleine sterretje' -- which have *neuter*

gender markers -- contribute to the token frequency of the *common* word 'ster'. That is, the de-word token counts reported in Table 2.3 contain an unknown number of occurrences that actually take het-word agreement in the corpus. As such, it may give a distorted picture of the relative salience of the two genders.

For these two reasons, I wanted to examine singular tokens only, and I wanted to make sure that singular diminutive occurrences do not inflate the de-word token counts. But what to do with the latter? Whereas CELEX takes the diminutive suffix to be an inflectional suffix, it can also be viewed as a derivational one (Geerts et al., 1984; Fromkin, Rodman and Neijt, 1986). Under a derivational interpretation, 'huisje' and 'sterretje' are both autonomous words. And they are het-words, regardless of the gender of their derivational root. Actually, the fact that the diminutive suffix can change the gender of a root word is probably the best argument for a derivational interpretation of this suffix (cf. Trommelen & Zonneveld, 1986). This means that diminutized de-words can justifiedly be counted as het-words.

The CELEX Dutch morphosyntactic word lexicon, which contains 399,816 morphosyntactic words, the 'inflectional variants' of 130,778 lemmas in the lemma
lexicon (Burnage, 1990), allowed for an analysis of singular nouns that would interpret the diminutive in this way. After again excluding the set of proper nouns in this database, I counted 164,297 common nouns, which correspond to 7,450,089 tokens in the INL corpus.[13] Within this set, there were 94,604 (58%) singular types, which corresponded to 5,741,040 (77%) singular corpus tokens. Of these 94,604 singular common noun types, only 4,211 (5%) turned out to be in the diminutive, against 90,393 (95%) non-diminutized types. And of the corresponding 5,741,040 singular common noun tokens, only 89,383 (2%) were in the diminutive, against 5,651,657 (98%) tokens that were not in the diminutive.

Table 2.4 shows the distribution of gender over 5,741,040 singular common noun tokens, with all diminutive tokens of de-words counted as het-word tokens. As can be seen, 67% of these tokens were unambiguously classified as de-word tokens, and 33% as het-word tokens, a de-het ratio of 2:1. This result, as well as its decomposition into three morphological classes, is actually very similar to the overall token count result reported in Table 2.3. The earlier overall result suggested that, of every noun encountered in running text, about every third turned out to be a het-

[13]Because this analysis starts out with the morphosyntactic variants of the common noun lemmas analyzed before, it should be no surprise that the *total* number of associated tokens is also the same as before. In the lexicon at hand, these 7,450,089 tokens have just been distributed over 164,297 morphosyntactic word types, instead of over the corresponding 92,628 lemma types.

word. Now we can see that the same holds forevery *singular* noun encountered in running text, even if we take diminutive de-word occurrences as het-words.
Taken together, the above analyses of the relative distribution of de- and het-words have shown that the earlier overall 3:1 dictionary estimates were correct, but also that the distribution of de- and het-word occurrences in running text is significantly less skewed. The running text estimate is roughly 2:1 in a global lemma token analysis, and exactly 2:1 in a singular noun token analysis that corrects for diminutive occurrences. Of course, the interpretation of this last result in terms of (implicit) gender salience rests on the assumption that singular occurrences of de- and het-words will on average be equally often marked for gender (e.g. by a definite article). This assumption cannot be checked with CELEX, but instead requires a full corpus analysis, which must remain beyond the scope of this thesis. In fact, the validity of a 'salience interpretation', or a 'storage interpretation', can in the end only be assessed against what we know about the gender-related processes in comprehension and production. As we begin to study these processes, though, we have to know something about the distribution of genders in the language environment of native speakers of Dutch.[14]

[14]With respect to this language environment, CELEX also allows us to look at some other aspects of how the gender system is embedded in the language. As mentioned before, several forms that realize particular morphemes in one or both of the two gender paradigms, notably 'het' and 'dat', also realize some completely different part of speech. But how often do they do this? In the 42 million word INL corpus, 'het' (and its reduced form "t') features 1,235,868 times, of which 867,947 (70%) are neuter singular definite article tokens, and 367,921 (30%) are expletive or personal pronoun tokens. The word-form 'dat' features 622,465 times, of which 263,104 (42%) are relative and demonstrative pronoun tokens, and 359,361 (58%) are complementizers. Thus, whereas 'het' is predominantly used as the neuter singular definite article, 'dat' will be used as a neuter pronoun in a minority of cases only (with an upper bound of 42%, since an unknown proportion of the demonstratives will be used independently, i.e., will not be reliably marked for gender).

	de	*het*	*rest*	*total*	*de-het ratio*
monomorphemic	1,828 k	795 k	1 k	2,625 k	2.3
	69.7%	30.3%	0.0%	(45.7%)	
complex	1,327 k	401 k	0 k	1,728 k	3.3
	76.8%	23.2%	0.0%	(30.1%)	
other	692 k	694 k	1 k	1,388 k	1.0
	49.9%	50.0%	0.1%	(24.2%)	
total	3,848 k	1,890 k	2 k	5,741 k	2.0
	67.0%	32.9%	0.0%		

Table 2.4 The distribution of gender over all *singular common noun tokens in the INL text corpus, with de-word diminutives counted as independent het-word tokens*. Shown are the number of de-word tokens (*de*), het-word tokens (*het*), and unclassifiable noun tokens (*rest*), each also as a percentage of the total number of tokens (*total*), as well as the number of de-word tokens divided by the number of het-word tokens (*de-het ratio*), within the monomorphemic noun stratum, the morphologically complex noun stratum, the remaining noun stratum, and the total noun stock (bracketed percentages express the relative size of a stratum within the total noun stock). All absolute token counts are in thousands (k = 1,000), and have been computed from the CELEX morphosyntactic word frequency counts. Any apparent inaccuracies in marginal percentages are the result of rounding.

Gender assignment

It is now time to examine a fundamental aspect of gender systems left untouched in my earlier description: how does a language with a gender system distribute its nouns over the available genders? Is this gender assignment a totally random affair, or is there some systematicity to it? I will first approach the issue at a purely descriptive linguistic level, i.e., as a question about regularity in language. The main reason for digging into this matter, however, is that it has important psycholinguistic implications. If the gender of a word is a totally random affair, then language users will simply have to memorize it in some way, along with the gender of tens of thousands of other words. To the extent that there is systematicity, though, language users may be able to exploit it. After having described the linguistic side of the issue, both in general and for Dutch, I will

therefore take a closer look at the logical implications for gender storage in the native speaker's mental lexicon.

Does gender assignment make any sense?

To the majority of linguists and psycholinguists, gender is an **essentially random** categorization, perhaps even the best example of arbitrariness in language structure (Zubin, 1992; Corbett, 1994). A frequently quoted passage from Bloomfield (1933) illustrates this idea:

> The gender categories of most Indo-European languages ... do no agree with anything in the practical world. ... There seems to be no practical criterion by which the gender of a noun in German, French, or Latin could be determined. (Bloomfield, 1933; quoted in Zubin & Köpcke, 1981, p. 439)

As another illustration, here is Maratsos (1979) characterizing the German gender system:

> The classification is arbitrary. No underlying rationale can be guessed at. The presence of such systems in a human cognitive system constitutes by itself excellent testimony to the occasional nonsensibleness of the species. Not only was this system devised by humans but generation after generation of children peacefully relearns it. (Maratsos, 1979; quoted in Zubin & Köpcke, 1981, p. 439)

Elsewhere, grammatical gender has been described as "an arbitrarily fixed characteristic of individual nouns" (Allerton, 1990, p. 94), and as usually operating along "seemingly arbitrary, even erratic lines". (Dekeyser, 1980; p. 97). So, although linguists have in the past often tried to interpret grammatical gender as a metaphorical extension of natural gender, it appears that they have now quite thoroughly embraced the idea that the two are just not systematically related, at least not across the entire stock of nouns in a language (Zubin, 1992).

But does this really mean that gender classifications are unsystematic? Several linguists have recently claimed that, although not as simple as once hoped for, gender classifications in fact have a large degree of systematicity. With respect to the allegedly nonsensible German gender system, for instance, Zubin and Köpcke (1981; see also Köpcke & Zubin, 1984; Zubin & Köpcke, 1986) have documented a number of phonological, morphological, and semantic assignment regularities, such as: "the more consonants a monosyllabic noun has in either initial or final position, the more likely it is to be masculine" (p. 440), "nouns forming plural with '-( )n' are feminine" (p. 443), and "nouns having extremely broad reference to objects having relevance to human needs are neuter" (p. 444).

Whereas Zubin and Köpcke were just concerned with German, Corbett (1991) studied the "assignment system" of dozens of gender languages around the world. On the basis of this survey, he arrived at the conclusion that gender assignment is **essentially systematic**, no matter what language you're looking at:

> Nouns may be assigned to genders according to semantic factors or according to a combination of semantic and formal (morphological and phonological) factors. While in some languages the rules are straightforward, in others they appear much less so. Nevertheless, in those languages which have been studied in depth, the gender of at least 85 per cent of the nouns can be predicted from information required independently in the lexicon. (Corbett, 1991, p. 68)

> Gender assignment is essentially systematic in all languages. (Corbett, 1994, p. 1350)

Corbett argues quite extensively that, although gender can always be derived from other properties of the noun at hand, languages differ in the kinds of properties they use. In a language such as Dyirbal, for example, gender is predominantly related to the **semantics of the referent**: words for male humans and non-human animates tend to have gender I, those for female humans, water, fire and fighting tend to have gender II, those for non-flesh food tend to have gender III, and the rest tends to have gender IV. In French, however, it is the **phonology of the noun** that matters most (cf. Tucker, Lambert & Rigault, 1977), whereas in Russian, gender assignment is to a large extent related to the **morphology of the noun**. And, as shown by Zubin and Köpcke, German appears to have a mixed assignment system, with gender being related to semantic, morphological, and phonological properties of nouns.[15]

[15]Whereas Corbett (1991) believes that systematicity rules supreme, Zubin and Köpcke (1981) actually entertained an interesting intermediate view. Faced with examples such as 'der Löffel', 'die Gabel', and 'das Messer' (spoon, fork, and knife), or 'der Hals', 'die Nase', and 'das Auge' (throat, nose, and eye), they pointed out that a language may strike a balance between motivated and arbitrary gender assignment because its native speakers have to deal with competing performance factors. Whereas limitations of memory and recall would push a gender system towards motivated assignment, the (anaphoric or deictic) referent tracking function of gender would push the system towards a specific form of arbitrariness: "The effectiveness of gender in this communicative function is increased if there is a *maximal differentiation* of gender among nouns referring to items that are likely to co-occur in the same perceptual field, or in the same text. This is precisely the case with nouns referring to parts of the face and head, and those referring to kitchen implements." (p. 447).

What about Dutch?

In general, then, whereas most linguists and psycholinguists view gender assignment as a largely random affair, an apparently well-informed minority takes it to be largely systematic. When it comes to Dutch, we find a similar distribution of views. Even though nobody denies the existence of some morphological and semantic regularity (e.g. all diminutives are het-words, most words for humans are de-words), the prevailing opinion is that Dutch gender is an essentially random affair (e.g. Geerts et al., 1984; Deutsch & Wijnen, 1985; Fontein & Pescher - ter Meer, 1985; Donaldson, 1987; de Houwer, 1987; Wijnen & Deutsch, 1987; Jescheniak, 1994). The two major reference grammars do in fact list an unexpected number of semantic, morphological and phonological regularities (Geerts et al., 1984, pp. 41-49; Donaldson, 1987, pp. 27-33), but they both emphasize the heuristic nature of the rules, pointing out that 'many' (Donaldson) or even 'most' (Geerts et al.) of the Dutch nouns remain beyond their scope.

Several linguists have tried to dig a little deeper, though. Some have made a case for substantial morphological conditioning of Dutch gender, and have confined chaos to the set of non-derived words (Trommelen & Zonneveld, 1986; van Beurden & Nijen - Twilhaar, 1992; Zonneveld, 1992). It has also been suggested that Dutch may instead be in the process of reorganizing its entire grammatical gender assignment around semantic principles (Fletcher, 1987). To my knowledge, however, only Frieke (1988) has looked for widespread systematicity in Dutch gender assignment. In an attempt to predict the gender of monosyllabic Dutch nouns, Frieke extracted 11 semantic and 3 morphological regularities from the literature (e.g. Geerts et al., 1984), and added 7 phonological regularities that emerged from a statistical analysis of 2943 monosyllabic nouns in the WNT (1954). Frieke's 'assignment system' included semantic 'rules' such as "nouns for very general things are more likely to be het-words", morphological rules such as "nouns that are nominalizations of other syntactic categories are more likely to be het-words", and phonological rules like "nouns with an initial consonant cluster containing an unvoiced plosive are more likely to be de-words". With 21 assignment rules in all, Frieke was able to correctly predict the gender of some 80% of the 763 monosyllabic nouns in a small test lexicon. After having adjusted his final estimate to "somewhere between 60 and 70 percent", Frieke therefore concluded that Dutch gender assignment is not arbitrary at all, and that grammars such as Geerts et al. (1984) and Donaldson (1987) simply underestimate the power of the regularities they mention.

Although Frieke's result is interesting, it should be qualified in a number of ways. First of all, the assignment system is for monosyllabic nouns only, and the rules have been tested on less than a 1000 of these nouns. It remains to be

seen, therefore, how well Frieke's assignment system would work when, for example, tested on the approximately 100,000 common nouns listed in the CELEX lemma lexicon.[16] Secondly, even if the assignment rules correctly predicted the gender of some 60 to 70 (or perhaps even 80) percent of the entire Dutch noun stock, one should bear in mind that, given a 3:1 lemma type distribution, a Dutch native speaker would be able to achieve the same degree of success by simply predicting that *every* word is a de-word. Viewed from this perspective, the reported coverage of 21 rules is somewhat disappointing.

Above all, though, we should ask about the meaning of a systematicity result such as Frieke's. For one thing, given an unlimited number of semantic, morphological, and phonological features (and their combinations) to play with, it would seem that one could always come up with a bunch of regularities. Unless we can restrict the set of possible predictors and their combinations in some principled way, there is no end to the gender assignment patterns we might find. Criteria of parsimony and elegance may help here, but they do not seem to be enough. Given that they make equally successful predictions, is a 100-rule assignment system better than a 200-rule one? Somehow, the most natural criterion for evaluating such systems is that of psychological reality. For example, one should ask whether the regularities are such that native speakers may plausibly (perhaps unconsciously) discover them. But the most important question should be this: do native speakers actually *exploit* the regularities captured in a particular assignment system?

Psycholinguistic implications

For those who see nothing but chaos in the assignment of gender, the above question simply does not arise. And, with nothing to be exploited, the psycholinguistic implication is clear -- native speakers must **store and retrieve** gender word by word:

[16]As Frieke pointed out, however, the gender of monosyllabic words may well be the most difficult one to predict. In the multisyllabic part of the Dutch lexicon, an assignment system could exploit many affix- and compound-related morphological regularities, as well as perhaps some stress-related ones.

> A person who has not studied German can form no idea of what a perplexing language it is. ... Every noun has a gender, and there is no sense or system in the distribution; so the gender of each must be learned separately and by heart. There is no other way. To do this, one has to have a memory like a memorandum book. (Mark Twain, 1879, *The awful German language*; quoted in Mills, 1986, p. 12)

If Dutch gender is essentially arbitrary, native speakers of the language *must* have something like a *Woordenlijst van de Nederlandse Taal* in their heads, with every word somehow marked for its gender (just common or neuter, of course).

For those who see extensive regularities in gender assignment, however, it is only natural to ask about exploitation. Indeed, if there are such regularities, they might be used by native speakers to **compute** the gender of words, and perhaps relieve those speakers from the necessity to learn every word's gender by heart. Rather than having a mental *Woordenlijst van de Nederlandse Taal*, native speakers of Dutch would then have something like Frieke's assignment rules in their head, using it to derive the gender of a noun from its other properties.

Corbett as well as Zubin and Köpcke have indeed wondered whether the regularity they see is actually being exploited by native speakers. Interestingly, though, they seem to have considerable difficulty imagining that it would *not* be. After some 60 pages on gender assignment systematicity around the world, for example, Corbett writes:

> We must ask what is the evidence for the psychological reality of the gender assignment systems discussed. The major evidence is, of course, the data already presented. Given the massive regularities established, and the ease with which native speakers use gender, the most plausible explanation is that speakers assign nouns to genders without difficulty simply by taking advantage of these regularities. ... Assignment rules are indeed part of the native speaker's competence, and not just regularities observed by linguists. (Corbett, 1991, p. 70)

Throughout the remainder of his text, Corbett makes it very clear that the 'taking advantage of these regularities' would be a routine thing. It is something that native speakers would do *as they speak*, whenever they need the gender of a noun, and regardless of how often they have used that noun before. And, because they would be able to do this, native speakers wouldn't bother to store gender word by word:

> The gender of the noun is normally predictable, on the basis of information which the speaker must in any case store in the lexicon. ... In this way we do not need to claim that gender languages are radically different from non-gender languages; they do not require an extra feature in the entry of each noun. (Corbett, 1991, p. 66)

Thus, the linguistic observation of extensive regularity in gender assignment is claimed to have two psycholinguistic implications: (1) native speakers use this

regularity to derive the gender for familiar words, as they speak, and (2) they therefore will not explicitly store gender in their mental lexicon. These are very strong, counter-intuitive, and interesting claims. If they are correct, we now know how gender is represented. But are they? What is the evidence for them?

Although Corbett takes the very *existence* of regularity to be the main evidence for on-line use by native speakers, he and others have several additional arguments:

1. Words borrowed from other languages acquire a gender, which shows that there is a mechanism for assigning and not just remembering gender. (Corbett, 1991, p. 7).

2. When presented with invented words, speakers give them a gender and they do so with a high degree of consistency. (Corbett, 1991, p. 7; see also Zubin & Köpcke, 1981).

3. Native speakers typically make few or no mistakes in the use of gender; if the gender of a noun were remembered individually, we would expect more errors. (Corbett, 1991, p. 7)
   On line recall of gender in speaking would be greatly hampered by intrusive errors, as it is in the speech of non-natives, if gender assignment were completely arbitrary. (Zubin & Köpcke, 1981, p. 447)

4. To have completely arbitrary gender assignment for the tens of thousands of nouns in the average educated speaker's lexicon would present an insurmountable task to the language learner. (Zubin & Köpcke, 1981, p. 447)

To what extent do these four arguments support the claim that native speakers compute the gender of familiar words, as they speak? First of all, I think that number 1 and 2 are simply not relevant to the issue. This is because, even if these two statements were correct, they are both about what I would call 'first-time' gender assignment. To the native speech community, lexical borrowings are new words. To the child or adult in a psycholinguistic experiment, invented words are also new words. The processes that guide the assignment of gender to *new* words need not have anything in common with the things that go on if a native speaker needs the gender of a word that he or she has used before. It is entirely possible that speakers exploit assignment regularity when they have to work out the gender for a new word, but at the same time simply retrieve the gender of the words they know already.

If anything, arguments 3 and 4 suggest a misunderstanding about what people can and cannot do. As for argument 3, neither Corbett nor Zubin and Köpcke give us any reasons *why* we would expect storage to lead to many more errors than computation. And it is not at all obvious why we should. One would not want to argue, for example, that the low incidence of word-form errors is evidence that speakers compute the form of a word from its meaning. If people can store the essentially arbitrary form of tens of thousands of words, why wouldn't they be able to store gender as well? Actually, given the weak reliability of most assignment rules proposed in the literature, and the fact that several conflicting rules may apply to a single word, one would rather expect *computation* to yield the highest error rate.

Argument 4 can be refuted in much the same way. Without giving any further information, Zubin and Köpcke claim that arbitrary gender simply cannot be learned. Clearly, though, the fact that children can acquire the arbitrary form of tens of thousands of words suggests otherwise. And, although systematicity in gender assignment will undoubtedly help the acquisition process, there is no a priori reason to expect that gender could not be learned without it. After all, there is systematicity in agreement (cf. Maratsos & Chalkley, 1980; Carstairs-McCarthy, 1994).

What about 'argument 0', the claim that the regularity *itself* suggests that native speakers compute the gender of every noun on-line, as they speak? And that, given they can do this, gender will no longer be stored? I would argue that regularity by itself doesn't suggest any of this. First of all, for each of the regularities that have been proposed, it is an empirical issue whether native speakers pick up on it. Secondly, even if they have, native speakers may not be able to use it fast enough to be of service as they speak. And thirdly, even if they would be able to do so, native speakers might store and retrieve the gender of known words just the same. In fact, they may not be able to avoid being redundant in their representation of such language facts. Memory doesn't seem to be a particularly expensive resource in the human system, and there is no reason why a single piece of knowledge cannot be represented in several different ways. Dutch native speakers are undoubtedly able to derive the gender of 'meisje', 'girl', from the fact that it is a diminutive, but *still* they may have stored it too.[17]

[17]The fact that a language user can represent linguistic knowledge redundantly may well surprise linguists such as Corbett, because they have been trained to represent linguistic knowledge as non-redundantly as possible. But, whereas linguists build *theories* of the mental lexicon, and are as such expected to be parsimoneous, a native speaker just builds a mental lexicon. Even if "the gender of the noun is normally predictable, on the basis of information which the speaker must *in any case* store in the lexicon (Corbett, 1991, p. 66; my emphasis), this speaker may not care and store the noun's gender just as well (see Lively et al., 1994, or Sandra, 1994, for a similar point).

A brief summary may be in order. Linguists such as Corbett, Zubin and Köpcke have not only argued that there is a lot of systematicity in the way languages assign their nouns to genders, but also that native speakers exploit this regularity. Specifically, native speakers would (1) compute the gender of familiar nouns on-line, as they speak, and they would (2) thereby avoid word-by-word memorization. In the above, I think I have shown that the arguments currently given to support these two hypotheses are flawed. Whatever the extent of regularities in how a language distributes its genders, we simply do not know yet whether native speakers make any use of it when they need the gender of words *they have used before*, nor whether this would keep them from explicitly storing the gender of those words in their mental lexicon.[18]

Still, even though the current arguments are wrong, the hypotheses may to some extent be right. If there is no regularity, gender must simply be stored word by word. But if there is, it *might* be used. Maybe not always, maybe not for all words, maybe in parallel to simple retrieval, but maybe. In this context, I think it is important to realize that the use of morphological regularities could have a special theoretical status that sets it apart from the use of semantics and phonology. Earlier, I mentioned that all Dutch diminutives take neuter gender agreement, e.g. 'het sterretje', 'het huisje', and that all noun-noun compounds take the agreements of the second noun, e.g. 'de veldslag', 'het slagveld'. In both cases, we could say that the gender of the result is simply 'inherited' from its rightmost constituent morpheme, i.e. its 'morphological head' (cf. Trommelen & Zonneveld, 1986; Scalise, 1994). If native speakers 'assemble' a morphologically complex word out of its constituent morphemes (see Feldman, 1995, for relevant theories), then the mechanism that causes *gender* to be inherited from the head may well also derive *other* features, such as word class, from the head. If that were the case, then it would be misleading to refer to this mechanism as being part of a 'gender assignment system'. The question would then be how people determine the *input* for such morphological inheritance, i.e. how they know the gender of the monomorphemic noun 'veld' in 'slagveld', or the gender of the diminutive suffix '-je' in 'huisje'.

[18]One might even argue that storage is, at some stage, a prerequisite for computation. Listeners in search for regularities cannot know in advance what features they should pay attention to. This makes it difficult to imagine a learning mechanism that would be able to induce rules without first storing a large set of exemplars. Of course, it is possible to imagine that the exemplars 'decay' once the assignment rules take over their job. Still, memorization will not have been avoided.

I think this explains why claims about morphological assignment (e.g. "native speakers of Dutch compute the gender of diminutives") seem a lot easier to accept than claims about semantic or phonological assignment (e.g. "native speakers of Dutch use the fact that flowers tend to be named with de-words"), at least when we are talking about speakers assigning gender to familiar words, as they speak. I wouldn't be too surprised if native speakers of Dutch would indeed turn out to compute the gender of a diminutive or a noun-noun compound on-line. But I would be very surprised if they would also turn out to routinely compute the gender of, say, a monomorphemic noun like 'veld'. Still, we cannot simply ignore the latter possibility. In view of the work by linguists such as Frieke (1988), Zubin and Köpcke (1981), and, above all, Corbett (1991), it would seem a bit rash to just continue to assume that monomorphemic gender is arbitrary, and therefore stored. It is up to empirical research, such as of the kind reported in Chapter 4, to decide the issue.

This concludes my linguistic exploration of grammatical gender. The remainder of this thesis is about the actual processing and representation of gender. I will begin by asking whether native speakers of Dutch use it as they try to recognize words.