Connecting conditionals: a corpus-based approach to conditional constructions in Dutch

Reuneker, A.

CHAPTER 4

Data selection and methodology

4.1 Introduction

In chapter 2, I argued for two conventional meanings of conditionals, namely their unassertiveness and its connectedness. I also argued the specification of these two meanings of conditionals into, for instance, uncertainty about p expressed in the antecedent, and a causal connection between p and q, to be conversational implicatures. In chapter 3, I reviewed existing classifications of types of conditionals, and from these accounts, linguistic features related to further implicatures of unassertiveness and connectedness were inventoried. In this chapter, I present the setup of the corpus study intended to address the relation between the aforementioned implicatures and grammatical features of conditionals.

The main aim of this chapter is to present the necessary preliminaries concerning the corpus study, so that detailed analyses can be provided in the next two chapters. We will discuss why a corpus study is an appropriate and promising methodology for the research questions presented in section 2.7. Although we have answered the first question in part, namely what specific implicatures are licensed through the conventional meanings of unassertiveness and the connectedness in conditionals, this was done solely based on existing and mostly theoretically motivated accounts. As the second research question specifically addresses the influence of grammatical features on these implicatures, it may seem the most direct and suiting approach to annotate (a selection of) types from these accounts in a corpus of natural language, together with the suggested grammatical features, and then assessing the predictive power of those

features for the types in each classification (see section 4.3). Such an approach, however, makes an important preliminary assumption: types (and features) must be reliably annotated for the analysis to succeed. Suggestions in the literature, however, indicate this is not self-evidently the case. Before any further steps were taken, therefore, an experiment was carried out to assess the reliability of applying the classifications of conditionals to actual language usage data. After reporting on this experiment, this chapter will lay the foundations for a corpus-based approach to the implicatures discussed in chapter 2. This paves the way for attempting to answer the second question, which concerns the extent to which the grammar conditional constructions licenses specific implicatures. Addressing methodological details will, of course, not address these questions directly, but it will guide the reader through some important preliminaries before the results in the next chapters can be presented and evaluated.

In section 4.2, I will present the results of an experiment in which the reliability of classification of conditionals in corpus data was evaluated. Then, in section 4.3, I will present arguments for a corpus-based study of conditionals in light of the framework of construction grammar, I motivate the current corpus-based approach, and I will discuss the focus on conditionals in a specific language (Dutch). In section 4.4, the data collection and the measures taken to arrive at a representative and balanced corpus of conditionals are presented. Next, in section 4.5, I will discuss the annotation of features and its reliability. In section 4.6, I will introduce the statistical procedures for the quantitative analyses applied to the data, of which the results will be reported in the next chapters. Finally, in section 4.7, I will draw a brief conclusion, before moving on to the next chapter, in which the distributions of the grammatical features of conditionals will be presented and discussed extensively.

4.2 Reliability of classification

4.2.1 Introduction

In various corpus studies on conditionals, existing top-down (deductive) classifications have been criticised for being too detached from actual language use (see e.g., Carter-Thomas & Rowley-Jolivet, 2008), or for being too dependent on contextual interpretation (see e.g., Ferguson, 2001). This criticism has led to several smaller-scale bottom-up (inductive) classifications, which better suit the data under investigation, but prohibit more general conclusions and replication. Claims such as the one by Dancygier and Sweetser (2005, p. 137) in which they remark that frequencies of different types of conditionals 'vary radically depending on the subject matter and the speaker's or author's goals' can only be tested properly if there is a reliable way of identifying such types in different datasets. On a related note, Verhagen (forthcoming) remarks that scholars analysing texts in detail, 'over and over again feel a need to define the categories anew, draw the boundaries somewhat differently than predecessors,

add other categories or distinguish subcategories [...]'. Whereas this passage may be read as a warning against the temptation of devising yet another classification of, in this case, conditionals, it also warns against the risks of applying theoretically motivated categories to language data, or applying categories constructed based on one data set onto another data set. This may indeed result in new categories and shifting boundaries, because such a deductive approach projects predefined categories onto the data (this issue is discussed extensively in the literature on framing analysis; see e.g., van Gorp, 2007, p. 72; Dirikx & Gelders, 2010, p. 733).

The aim of this section is to address the reliability of annotation of types of conditionals in natural language corpora as a preliminary for further steps in this study. In this section, therefore, I discuss an experiment reported on by Reuneker (2017a) in which the reliability of applying three classifications discussed in the previous chapter was critically assessed. Next to presenting the experiment and its results, I will discuss the implications not only for this study, but also for future research involving the classification of natural-language data.[1]

In section 4.2.2, I will discuss the evaluation of inter-annotator reliability, focusing on corpus studies of conditionals and related topics. Next, in section 4.2.3, I will present the data and method used in the experiment, and in section 4.2.4, the results be presented. In section 4.2.5 I will draw conclusions, before moving on to the corpus setup for the subsequent steps in this study.

4.2.2 Evaluating reliability

As a number of authors note, the application of classifications to natural (language) data is not only a time-consuming and challenging, but also an important measure of its validity (see e.g., Artstein & Poesio, 2008, p. 557; Bolognesi, Pilgram & van den Heerik, 2017, pp. 1985, 1988).[2] As a preliminary test for further data analysis in this study, Reuneker (2017a) therefore subjected the classifications by Quirk et al. (1985), Dancygier and Sweetser (2005), and Athanasiadou and Dirven (1997a) to an experiment on annotation reliability.[3]

Although Athanasiadou and Dirven present their classification in terms of prototype theory, as does Dancygier (1998), they ultimately classify each conditional sentence as one type. Sweetser (1990, pp. 124–125) explicitly mentions

[1] Next to the question concerning the reliability of classifications, the study by Reuneker (2017a) serves as a methodological case study for comparing reliability measures between different classifications by introducing ways of in-depth comparison based on combinatorial agreement-distributions. These issues will largely be ignored here due to restrictions of space.

[2] See also Levshina and Degand (2017, p. 146), who propose to deal with the 'high cost of manual annotation of discourse connectives' by using automatic annotation of lower-level ('semantic and syntactic') features for pre-annotation of coherence relations, after which these annotation should be verified and corrected by manual analysis.

[3] For details on these classifications, see sections 3.3.4, 3.3.7 and 3.3.9 respectively. For reasons of space, the overviews of these classifications will not be repeated here.

the problem of ambiguity for her analysis of conditionals in the content, epistemic and speech-act domain (see section 3.3.7): 'A given example may be ambiguous between interpretations in two different domains, [...], but no one interpretation of an *if-then* sentence [...] simultaneously expresses conditionality in more than one domain'. This shows that the authors implicitly strive for mutually exclusive types, contrary to prototype categories, which can have 'fuzzy boundaries' (cf. Taylor, 2003, p. 51). In the previous chapters, we also saw numerous examples of ambiguity between, for instance, specific and general conditionals, past tense marking temporal or epistemic distance and problems alike. This means that, in annotation, in such cases a choice must be made, because the form of an utterance does not fully determine the intended meaning.

As we saw in chapter 3, various classifications of the same phenomenon are offered in the literature. Although the terminology differs, in a number of cases, these classifications classify conditionals in a highly similar way. Comparing the classifications reveals, however, that there is no one-to-one relation between the types and sub-types in the various accounts. Whereas the example in (1) would be consistently classified as an indirect, pragmatic or speech-act conditional in the classifications by Quirk et al., Dancygier and Sweetser, and Athanasiadou and Dirven respectively, an example such as in (2) would not.

(1) So: if you're interested and you don't have any plans yet, the Dutch Philharmonic Orchestra plays Tchaikovsky tonight.

(2) If that's art, then I'm an artist too!

Quirk et al. (1985, p. 1094) distinguish a rhetorical type of conditional for the example in (2), whereas Athanasiadou and Dirven (1997a) would classify this example as a subtype of pragmatic conditionals. As Quirk et al. (1985) place rhetorical conditionals outside their direct-indirect distinction, and pragmatic conditionals would fall inside the indirect class, this amounts to an inconsistency between classifications. Dancygier and Sweetser (2005) do not analyse rhetorical conditionals as a separate type, but the example satisfies the criteria of epistemic conditionals, because the falsity of the antecedent licenses the conclusion in the consequent, albeit indirect through the projection of falsity from the consequent. Epistemic conditionals are a sub-type of non-predictive conditionals, however, while they are direct conditionals in Quirk et al.'s (1985, p. 1091) account and a subtype of either course-of-event or pragmatic conditionals in Athanasiadou and Dirven's account. Such discrepancies between classifications are, in themselves, not problematic. As long as classifications are viewed as artificial constructs rather than reflections of natural systems (Sandri, 1969, pp. 86–87), different perspectives and organisations can co-exist. Although this view shifts the question from 'Which classification is right?' to 'Which classification is able to explain the data best and most efficiently?', preliminary to both questions, however, is the question of reliability: 'To what extent are raters able to apply classifications consistently to real data?'.

Spooren and Degand (2010, p. 242) remark that 'there is presently no tradition in the field of corpus-based discourse studies to report agreement measures', which may, in part, be due to low agreement scores reported in studies that do (see also Mulken & Schellens, 2012, p. 43; Neuendorf, 2017, chapter 6). While recent research on, for instance, coherence relations does show an increasing number of studies explicitly addressing the question of inter-annotator agreement (see e.g., Rehbein, Scholman & Demberg, 2016; Bolognesi, Pilgram & van den Heerik, 2017; Prasad et al., 2017; Levshina & Degand, 2017; Hoek, 2018; Hoek, Evers-Vermeul & Sanders, 2019), Spooren and Degand's remark clearly applies to the literature on conditionals. In discussing their corpus annotation, Renmans and van Belle (2003, p. 152) remark that 'obviously, there are still no reliable, let alone objective ways to identify the underlying semantico-pragmatic reading of a certain conditional sentence, the classification was to a large extent based on personal interpretation and accordingly, could have been subject to human error'. Most studies are not as explicit on this issue, however. Athanasiadou and Dirven, for instance, provide frequencies of attested types, but do not mention how these results were obtained and whether or not the annotations were evaluated in terms of reliability. Dancygier and Sweetser use examples from corpora, but no frequencies, nor reliability measures are provided. Reliability is, however, a prerequisite for the demonstration of validity of a classification scheme, i.e., showing 'that the coding scheme captures the "truth" of the phenomenon being studied' (Artstein & Poesio, 2008, p. 557). Low reliability scores signal a problem, as they indicate that 'the theoretical categories cannot be applied with any confidence' (Spooren, 2004), and that types in classifications are 'vague, in the sense that categorisations are non-replicable, and consequently unfit as a basis for theory building' (Spooren & Degand, 2010, p. 242).

Contrary to its relative absence from the literature on conditionals, the issue of reliability is of major importance to the study of conditionals, as the assignment of specific uses to classes of conditionals is, inevitably, based (at least partly) on interpretation. Add to this the observations by Miltsakaki et al. (2004) and (Prasad et al., 2008; see also Hoek, Evers-Vermeul and Sanders, 2019, p. 19) that the annotation of coherence relations marked by explicit connectives results in higher agreement scores than the annotation of implicit relations, and it is clear that the notion of reliability is vital for the study of conditionals in corpora, as it allows for the assessment of the extent to which classification results are 'independent of the measuring event, instrument or person' (Kaplan & Goldsen, 1965, p. 83). This is especially relevant for the annotation of types of conditionals in this study, as they are analysed in terms of conversational implicatures (see chapter 2), and as such, they are non-conventional and not or only partly marked for the type of unassertiveness and connectedness. Reliability is understood in this study as the combination of stability (do rater's judgments remain constant over time?) and replicability (can judgments be reproduced among raters?). As such, it differs from measures of validity, which represent the 'the extent to which [both] raters classify sub-

jects into their true category' (Gwet, 2014, p. 314). In the experiment reported on by Reuneker (2017a), both stability and reliability were investigated. We will turn to the data and method briefly in section 4.2.3 below, before moving on to the results and their implications in section 4.2.4.

4.2.3 Data and method

To measure the reliability of applying the aforementioned classifications to natural-language data, an experiment was conducted in which a group of trained students (henceforth: *raters*) classified a set of conditionals from the CONDIV corpus of written Dutch (Deygers et al., 2000) and the CGN corpus of spoken Dutch (Oostdijk, 2000). The experiment followed a within-participants design to control for effects of individual differences in linguistic knowledge and understanding of the materials. The raters were 27 native speakers of Dutch, and students of Linguistics at Leiden University (22 female, 5 male) with an average age of 22.7 years (sd=5.1). The raters participated for course credit in a course on corpus linguistics and classification of conditionals. For each classification, the original article or chapter was distributed as part of the course materials. Raters were asked to read the text and classify a set of conditionals accordingly prior to the class in which the classification was discussed. Both the examples provided by the authors and real usage data were used as training material. Examples and counter-examples of types were discussed collectively. A week before the experiment, raters were presented with an overview of the classifications, including criteria for each type (see Appendix E), in order to enable them to evaluate their understanding of the source texts and familiarise themselves with the instructions for the experiment.

The items were Dutch conditional sentences and consisted of 3 practice items to familiarise raters with the task, 14 items from the written corpus, 9 items from the spoken corpus, 8 control items, which were variations on examples from the literature, and 2 test-retest items (for these materials, see Appendix E). All items included one sentence preceding and one sentence following the conditional sentence. The conditional sentence itself was presented in bold. Each rater annotated 33 items according to the three classifications mentioned above. In order to control for memory and practice effects, the order of classifications applied was counterbalanced using a latin-square design. Within each block, the conditionals were presented in random order. Per item, raters chose a type, indicated their confidence on a 5-point Likert-scale, and optionally included a comment. In total, each rater classified 102 sentences.

4.2.4 Results

The first step was to select those raters who were able to correctly apply the classifications. To do so, eight control items were randomly presented in each of the three trials. These control items were based on the aforementioned criteria and the examples provided by the authors of the classifications, and could be

called 'idealised examples'. No authentic examples by the respective authors were included to avoid memory effects. As the goal of the experiment was to measure the reliability of existing classifications when applied to natural-language data, it was found necessary to control for confounding factors related to participant's individual abilities. In short, the control items allowed for the qualification of only those raters who were able to correctly classify idealised examples. For the selection procedure described here, a gold standard was available, because the items were specifically designed to belong to specific classes of the classifications. Therefore, not reliability, but validity was calculated for each participant (i.e., how well a rater's classification judgments confirm to actual values; also called *accuracy*). Validity was calculated by dividing the number of true positives (correct answers) by the total number of classifications made. The results are presented in Table 4.1 below.

Table 4.1:
Validity for control items per classification (before selection, N=27)

	Validity	
Classification	*mean*	*sd*
Quirk et al. (1985)	0.84	0.12
Athanasiadou and Dirven (1997a)	0.81	0.17
Dancygier and Sweetser (2005)	0.68	0.16

Instead of using an arbitrary cut-off point or often criticised guidelines for agreement scores such as those offered by Landis and Koch (1977), negative deviation from the mean validity was used. If a rater's accuracy score was more than one standard deviation lower than the mean (a z-score of -1 or less), this was taken to signal an inability to classify idealised examples and, thus, an inadequate understanding of the task. Nine raters were excluded from further analysis. As can be seen in table 4.2, this resulted in higher accuracies and lower deviations.

Table 4.2:
Validity for control items per classification (after selection, N=18)

	Validity	
Classification	*mean*	*sd*
Quirk et al.	0.89	0.06
Athanasiadou and Dirven	0.83	0.15
Dancygier and Sweetser	0.75	0.13

Both Table 4.1 and Table 4.2 suggest a difference in validity of ratings between Quirk et al.'s and Athanasiadou and Dirven's classification on the one hand, and Dancygier and Sweetser's classifications on the other. A repeated-measures ANOVA ($F(2,36)=7.58$, $p=0.0018$) confirmed that classification as a factor had a significant effect on accuracy within each participant. A post-hoc test using Bonferroni correction showed that the validity of Dancygier and Sweetser's classification (0.75) differed significantly ($p<0.001$) from those by Quirk et al. (0.89) and Athanasiadou and Dirven (0.83). This shows that participants had more difficulty classifying idealised examples of conditionals when using Dancygier and Sweetser's classification than when using those of Quirk et al. and Athanasiadou and Dirven.

In contrast to the measurement of validity, no gold standard was available for the corpus data, i.e., actual language data do not come with a 'correct label'. Therefore, agreement coefficients in the form of Krippendorff's Alpha (Krippendorff, 2004; Hayes & Krippendorff, 2007) were calculated. Results are presented in table 4.3 below. Note that scores are provided for agreement on the level of main types, and on the level of sub-types, and only for the 18 raters selected in the previous procedure.

Table 4.3:
Agreement for control and corpus items

	Control		Corpus	
Classification	main type	sub-type	main type	sub-type
Quirk et al.	0.87	0.69	0.53	0.41
Athanasiadou and Dirven	0.59	0.45	0.31	0.29
Dancygier and Sweetser	0.55	0.56	0.32	0.28

Note. Agreement scores for both main types and sub-types are reported in terms of Krippendorff's Alpha.

What this table shows, is that the agreement between the 18 raters on corpus items is consistently lower than their agreement on control items.[4] This shows that judgements were indeed more reliable for idealised examples than for real,

[4]The small and reversed difference between main level and sub-level control items for Dancygier and Sweetser's classification can be explained by the small difference between the occurrence of four categories in the results for main types, and only five categories in the results for sub-types. The reason that this account is not brought down to two categories (i.e., predictive or content and non-predictive) is that studies using Dancygier and Sweetser's classification distinguish mainly between content, epistemic, speech-act and metalinguistic conditionals, not between predictive and non-predictive conditionals. Further note that Krippendorf's Alpha corrects for the number of categories (see also section 4.5).

attested conditionals.[5] The agreement on control items ranges between 0.55 and 0.87 and is higher than the agreement on corpus items, which ranges from 0.31 to 0.53. Both on control items and on corpus items, Quirk et al.'s classification results in substantially higher agreement scores than Athanasiadou and Dirven's and Dancygier and Sweetser's. Although these latter two scores are low already, the corpus scores are lower still. What can also be seen, is that, when sub-types are taken into account, the reliability decreases, which is consistent with other observations in the literature (see e.g., Spooren & Degand, 2010; Bolognesi, Pilgram & van den Heerik, 2017, pp. 1993–1994). In the results presented and discussed below, only main types are taken into account.

To allow for a more detailed analysis, a distribution of agreement coefficients was calculated. For all combinations of raters, Krippendorff's Alpha was calculated, resulting in 153 coefficients per classification. A repeated-measures ANOVA showed that the independent variable *classification* had a significant effect on the dependent variable *agreement* ($F(2,453)=37.43$, $p<0.001$). A post-hoc test using Bonferroni correction confirms what Table 4.3 already suggests, namely that the significant difference lies between Quirk et al. (1985) on the one hand and Athanasiadou and Dirven (1997a) and Dancygier and Sweetser (2005) on the other. In other words, raters were more reliable in their application of Quirk et al.'s classification, than in their application of the other two classifications.

Whereas inter-rater reliability is concerned with the agreement between different raters, intra-rater reliability is concerned with the 'self-reproducibility' (Gwet, 2014, p. 200) or 'stability' of classifications, which is also called 'test-retest reliability'. Krippendorff (2004, p. 215) argues that intra-rater reliability is a far weaker measurement of reliability than inter-rater reliability, because it only measures the degree to which classification results can be replicated by one rater, instead of by different raters. However, as, for instance, Verhagen and Mos (2016, p. 336) argue, the processing of linguistic material of an individual may vary between moments, which calls for the measurement of 'individual variation and its underlying dynamics' (see also Dąbrowska, 2014). For a full inquiry into the stability of the application of classifications to natural-language conditionals, the calculation of intra-rater reliability should be based on the same rationale as that of inter-rater reliability, i.e., classifying one item several times into the same class may reflect consistency, but can also be the result of chance (see e.g., Gwet, 2008). As this study's main focus is on inter-rater reliability, the number of test re-test items and the number of iterations was limited to keep the task manageable for raters. Consequently, only percentages of intra-rater agreement per classification could be calculated.[6] For each classi-

[5] Note the difference between Table 4.1 and Table 4.2, and Table 4.3. In tables 4.1 and 4.2, validity scores are presented, in which no correction for chance agreement is performed, while in Table 4.3, this correction is applied, which results in lower scores due to the distribution of categories and answers.

[6] Keeping in mind the earlier remarks on the use of percentages (see above), these figures must be interpreted with caution.

fication, one item from the spoken corpus and one from the written corpus was adapted to function as a test-retest pair. To rule out possible confounding variables, care was taken to apply changes only on the lexical-semantic level of the utterance, while keeping their syntactic structures constant (see the materials in Appendix E). The results are presented in table 4.4 below. The percentages suggest that rater's judgments are stable, and the fact that the intra-rater reliability scores are high suggests that the low inter-rater agreement scores are not the result of random assignment of conditionals to types.

Table 4.4:
Intra-rater reliability on corpus items

Classification	Agreement
Quirk et al.	91.7
Athanasiadou and Dirven	77.8
Dancygier and Sweetser	91.7

Note. Agreement scores are reported in terms of raw agreement (i.e., percentages).

In addition to the annotation of types of conditionals, raters also reported their confidence in the type chosen on a 5-point Likert-scale (1=very uncertain; 5=very certain). There proved to be a correlation between inter-rater agreement and confidence (*Pearson's* $r(16)=0.73$, $p<0.05$), meaning that items that reached low agreement (closer to 0.0 agreement) were found harder to classify by raters (closer to 1 on the confidence scale). This suggests that raters were aware that certain items were harder to classify than others.

While the data presented so far allow for a straight-forward comparison of agreement scores, they do not provide a detailed picture of agreement on item level (i.e., per conditional), because the agreement scores compress a multitude of ratings into a single figure. These results, therefore, cannot be used directly for more detailed analyses, such as an analysis of variance or within-item agreement. Consequently, two different steps were taken. First, a pair-wise combinatorial distribution of Krippendorff's Alpha coefficients was generated. This is, however, not a trivial task, as agreement coefficients are normally calculated over the distribution of items and raters, not over the distribution of ratings per item. Therefore, the average agreement per item (O'Connell & Dobson, 1984; Schouten, 1982) was calculated.[7] The results are presented in Appendix E, and show that the majority of corpus items score in the range of a 'slight' (<0.20) to 'fair' (0.21-0.40) level of agreement (cf. Landis & Koch, 1977). A large number of items turned out to be problematic. The distribution presented in the aforementioned appendix allowed for the identification of the most

[7]For an R package capable of estimating O'Connell-Dobson-Schouten coefficients, see https://github.com/mclements/magree.

problematic cases in each individual classification, and in general. Here, I will only discuss briefly one the most problematic cases for each classification. For a more elaborate analysis, see Reuneker (2017a).

In case of Quirk et al.'s classification, raters agreed only very weakly on the conditional in (3).

(3) We moeten oppassen dat de toeloop op de opleidingen in Limburg niet te groot wordt. Het is gevaarlijk als 'genoeg werk' het enige argument is om aan de Pabo te gaan studeren. (limburg/nieuws04)
We must be careful that the number of students for the study programs in Limburg does not grow too large. It is dangerous if 'enough work' is the only argument to study for teacher.[8]

While most raters decided to annotate the conditional in (3) as a direct conditional (66.7%) and 5.6% as an utterance conditional, 27.8% chose not to classify this item. This could be due to the *als* 'if' clause functioning as a subject to the evaluation in the matrix clause ('it is dangerous'). A relation between antecedent and consequent as a subject that is evaluated is not present in any of the classifications, and it could be the case that this 'evaluative conditional' is a language specific construction, although Ford and Thompson's (1986, p. 368) results suggests otherwise, as they show this use is also present in English and even quite frequent in case of sentence-final antecedents (see sections 5.2 and 5.6). In case of Dancygier and Sweetser's classification, raters also agreed weakly on the conditional in (3). 27.8% of the raters classified (3) as a predictive conditional, another 27.8% as a speech-act conditional, 22.2% as an epistemic conditional, 16.7% as a meta-linguistic conditional, and 5.6% did not classify the conditional. Using Quirk et al.'s classification scheme, this conditional resulted in problems as well, but raters were somewhat more unanimous. It is unclear why this should be the case, as the reasons for the most likely candidate of utterance conditional also apply to speech-act conditionals. As a relatively small group chose not to annotate this example, there must be another reason for the scattered distribution. What could be the case, is that the main parameter of backshift to distinguish between predictive and non-predictive conditionals in Dancygier and Sweetser's classification (see section 3.3.7) led raters to choose the predictive type, as verb tense in Dutch might be a less reliable source of conditional relation than is the case in English, which is indeed what we will test in section 5.4 and chapter 6. The group of raters choosing the epistemic type may have done so by interpreting the consequent as a conclusion, consequently viewing the antecedent as an argument.

Another low score was obtained for the conditional in (4) below.

[8]Examples in this section are taken from the Condiv Corpus of written Dutch (Deygers et al., 2000) and from the 'Corpus Gesproken Nederlands' (CGN; Oostdijk, 2000). See Appendix E for details.

(4) Mmm? Als je 't niet zou weten dan hoor je niet dat de radio aan staat. nee, maar was trouwens wel gaaf dat concert. (fn000411)
Mmm? If you wouldn't know then you would not hear the radio is turned on. No, but the concert was really cool by the way.

For (4), 44.4% chose the predictive type, which seems the right choice, given the hypothetical backshift in the antecedent adding to a counterfactual interpretation. However, there were also raters annotating this example as a speech-act conditional (11.1%), an epistemic conditional (27.8%), and a meta-linguistic conditional (5.6%). 11.1% chose not to annotate this example. The high percentage of raters opting for the epistemic type may be due to the fact that the antecedent concerns knowledge ('if you wouldn't know') and might therefore be easily interpreted as an argument for a conclusion in the consequent. A related indication found in the distribution of item-agreement scores concerns the distinction between the direct and indirect types in Quirk et al., 1985's classification on the one hand, and the predictive and non-predictive types in Dancygier and Sweetser's classification on the other hand. Although there are differences, in many cases these distinctions should result in the same outcome, but raters were able to apply Quirk et al., 1985's distinction more reliably than Dancygier and Sweetser's distinction. This discrepancy seems to be connected to the aforementioned problems in distinguishing between predictive and epistemic conditionals, which in Quirk et al., 1985's classification are both considered direct conditionals. Finally, in case of Athanasiadou and Dirven's classification, raters agreed only weakly on the example in (5).

(5) Maar dat kan niet want de ZCTU beschikt niet over de kwaliteiten van een president, aldus Moegabe, die er voor de goede orde aan toevoegde: "De vakbonden vergissen zich als ze geloven dat ze sterker zijn dan mijn regering. Ik waarschuw de ZCTU. Ik maak geen grapjes, ik ben bloedserieus." (tele/nie_s5)
But that is not possible, because the ZCTU does not have the qualities of a president, said Mugabe, who, for the record, added: "The unions are mistaken if they believe they are stronger than my government. I warn the ZCTU. I'm not kidding, I'm dead serious."

The conditional in (5) was annotated as a hypothetical conditional by 44.4% of the raters, as a pragmatic conditional by 27.8%, and as a course-of-event conditional by 27.8%. This example indeed does not fit easily into one of the types of Athanasiadou and Dirven's classification, which is also reflected in low certainty scores provided by the raters. It can be viewed as a hypothetical conditional, as the situation in the antecedent presents a specific hypothetical state of affairs. The consequent however presents an evaluation of the situation in the antecedent, just as in (3). If the evaluation is seen as a conclusion based on an argument, it would amount to a pragmatic conditional of the inferential sub-type.

Reliability is a prerequisite for the demonstration of validity of a classification scheme, i.e., showing 'that the coding scheme captures the "truth" of the phenomenon being studied' (Artstein & Poesio, 2008, p. 557). The analysis of problematic cases above, i.e., those cases for which reliability was lowest, provides suggestions for possible 'blind spots' of classification schemes. As Carter-Thomas and Rowley-Jolivet (2008) suggest, classifications may be too idealised and detached from actual language use, possibly because the selected examples are not representative of all corpus data (cf. the principle of 'total accountability'; see McEnery and Hardie, 2012, pp. 14–18). Furthermore, it might be the case that the criteria offered by the respective authors are not clear enough to be applied to other data. In this sense, the classification of implicatures of connectedness is comparable to the annotation of (other) coherence relations, which are also, to a certain extent, interpretative, rather than determined by grammatical features (cf. Spooren, 2004; Sanders & Spooren, 2007; Spooren & Degand, 2010; Artstein & Poesio, 2008).

A possible source of low validity and reliability may be the use of Dutch corpora, while the classifications under inspection are based on and targeted at English. This may lead to problems when encountering language specific types of conditional relations. For example, the scalar type of Quirk et al.'s rhetorical conditional in (6) below can be expressed using a conditional in Dutch, in which case the antecedent needs to be altered to great extent, as in (7). In Dutch, this meaning is expressed more frequently by other means than a conditional. It can be expressed by a rhetorical conditional, as in (8), which also needs to be altered to great extent, not in the least by exchanging propositions between antecedent and consequent.

(6) The package weighed ten pounds if it weighted an ounce. ['The package certainly weighed ten pounds.'] (Quirk et al., 1985, p. 1095)

(7) Het pakketje woog (zeker) tien pond, als het (al) niet meer was.
The package (certainly) weighed ten pounds, if it wasn't (even) more.

(8) Als het pakketje geen tien pond woog, dan eet ik mijn hoed op.
If the package did not weigh ten pounds, then I will eat my hat.

Conditionals in which the *if*-clause functions as the subject of an evaluation in the main clause, which are common in Dutch, resulted in low agreement scores too. This may be due to their absence from or only brief mentions in classifications of English conditionals. Comparisons between, in this case, Dutch and English form a testing ground for the applicability of classifications to other languages. It provides, as Verhagen (2007, p. 272) argues, 'the insight that grammars do not only consist of regularities on the one hand, and idiosyncracies on the other. Rather, some combination of the two seems to be the rule rather than the exception (paradoxically so), so that the balance is always an open issue, and thus deserves investigation'. The current results show that a comparative perspective may help understand this balance between regularities and idiosyncracies in conditional constructions. Furthermore, it has made a

case, as mentioned earlier, for not investing in yet another classification with (sometimes slightly) differing types and boundaries (see also sections 4.4 and 4.5, and Verhagen, forthcoming), but in testing an important aspect of existing accounts on real language data: their applicability and reliability.

4.2.5 Conclusion

This section reported on an experiment in which the reliability of applying three classifications on corpus data was evaluated. This was done both as a preliminary quality measure for subsequent steps in the current study, as to make a case for the application of reliability measures in corpus studies on conditionals.

Three classifications were compared in terms of validity and reliability with respect to both idealised examples and corpus data. While raters were able to apply the classifications to the former, they were unable to classify conditionals from actual corpus data with a sufficient level of reliability. In other words, annotations were not sufficiently replicable between raters. The results of this experiment suggest that replication and generalisation may be compromised by the reliability of classifications. Low reliability scores for the classification of conditionals may be the result of a number of problems, and it is insightful to apply Spooren and Degand's (2010) distinction between two types of disagreement. First, disagreement can be the result of simple coding errors, which we will encounter and deal with in the next chapter. Second, as language underspecifies meaning and context guides interpretation, there is ambiguity as a result of linguistic underspecification, which, Spooren and Degand argue, puts 'perfect agreement' out of reach. The second type of disagreement should however tell us 'something about the stability of our coding scheme and the theoretical conclusions that can be drawn from our analysis' (Spooren & Degand, 2010, p. 251). The low agreement scores reported in this section thus raise the question to what extent existing classifications can be applied to actual language use. This experiment may therefore be seen as a methodological contribution to the study of conditionals, and to corpus linguistics in general, because novel ways of comparing agreement distributions were presented, including item-wise agreement computations in order to identify problematic cases. The methodology may be useful to identify items that resist classification, and may subsequently be used to improve classification schemes by specifically addressing these issues. One step that would have improved the present experiment was a group discussion after the classification task, as the disagreements among raters were not discussed, prohibiting the identification of reasons for disagreement and types of disagreement. On the other hand, reliable classification should, ideally, be the product of independent classification.

The results of this experiment support the analysis of the types of conditionals discussed in section 3.3 as conversational implicatures of connectedness, and it raises the question to what extent they are generalised. Grammatical form may support and license, but will not fully determine the type of connection

between antecedent and consequent, as is also suggested throughout the literature discussed in chapter 3. The results of this experiment show that the application of classifications to conditionals in corpora yields low reliability, which has important ramifications for the available data analyses in answering the research questions of this study. In the next section, therefore, I will discuss the data-analytic approach that will be used in the remainder of this dissertation.

4.3 A corpus-based approach to conditional constructions

4.3.1 Introduction

The central question in this dissertation boils down to the following questions: Which implicatures are licensed by means of conditionals, and how does their grammatical form support these implicatures? These conversational implicatures were narrowed down to further specifications of the conventional meaning aspects of unassertiveness and connectedness. In chapter 3 we discussed types of conditionals in these terms, with additional focus on the grammatical features they were related to in the literature. In the previous section, however, we saw that the annotation of such types in corpora yielded problems with respect to reliability.

The aim of this section is to address this issue, and to deal with its ramifications for the remainder of this study. In section 4.3.2, I will first address the reasons for employing a usage-based approach, and more specifically, the reasons for a corpus-based approach to conditionals in the Dutch language. Next, in section 4.3.3, I will provide arguments for a bottom-up approach to data analysis, which, as will be discussed in section 4.3.4, will lead to the choice for a cluster analysis of conditionals. After drawing brief and intermediate conclusions in section 4.3.5, I will continue by presenting the corpus setup in section 4.4.

4.3.2 Constructions and corpora

In order to analyse the form and meaning of conditionals in unison, I will analyse them in terms of construction grammar, (cf. Goldberg, 1995), a framework we previously discussed in section 2.2. The reason for adopting a usage-based approach in this study is that the meaning aspects mentioned are fundamentally 'aspects of the use that human beings make of language' (cf. Verhagen, 2005, p. 24). This also allows for acknowledging that some aspects of this use have become conventionalised by strong generalisations over 'usage events' by speakers and hearers, whereas other, less generalised aspects are, by definition, more contextual and more appropriately described in terms of conversational implicatures.

Within construction grammar, it is customary to investigate to what extent the formal (i.e., grammatical) features of an utterance contribute to its meaning (see e.g., Bybee, 2013, p. 51; Goldberg, 1995, pp. 1–9; Goldberg, 2019, pp. 2–3), while, at the same time, leaving room for idiomaticity, i.e., the idea that a construction may 'specify a semantics (and/or pragmatics) that is distinct from what might be calculated from what might be calculated from the associated semantics of the set of smaller constructions that could be used to build the same morphosyntactic object' (Fillmore, Kay & O'Connor, 1988, p. 501). With respect to conditional constructions, Dancygier and Sweetser (2000, p. 138) argue that 'what is needed is an analysis which uses parameters of constructional meaning (verb forms, clause order, intonation, use of mental space builders) to outline the range of constructions which participate in the construal of related meanings (causality, sequentiality, conditionality), and explores the similarities and differences between the constructions with respect to these parameters'. In my view, especially because we discuss part of the non-truth-conditional meaning of conditionals in terms of implicatures, the way to do this, is to conduct a corpus-based study of the distributions of these 'parameters' or linguistic features. In chapter 5 these will be systematically investigated and discussed, but before addressing the need to carefully construct a representative collection of conditionals in section 4.4, we will discuss the choice of a language-specific study.

As we saw in chapter 3, most accounts of conditionals are based on English, with the exception of the classic accounts of conditionals in Ancient Greek. As I aim at analysing both the form and meaning of conditionals, and especially their connection, it is needed to construct a language-specific corpus in order to provide a detailed analysis. The reason for this, in line with Croft's (2001) arguments for his 'Radical Construction Grammar', is that it is not to be expected that the systematic relations between form and meaning in English conditionals, such as *will* in consequents to mark q as the causal consequence of p in the antecedent, will be universals. Rather, it is to be expected that a language, and, more to the point, conditionals in a specific language, will have meanings that depend on the part-whole relation between the conditional construction and its elements, rather than meanings that can be derived from independently definable, language independent meanings of its (grammatical) elements. In other words, we need to 'account for the diversity of the syntactic facts of a single language as well as the syntactic diversity of the world's languages' (Croft, 2001, p. 3). This is reminiscent of Verhagen's (forthcoming) observations of differences in how speech and thought are construed in different languages, even closely related languages. He shows how, in research, the speech and thought representation (STR) categories of 'direct discourse', 'indirect discourse' and 'free indirect discourse' are used as 'relatively abstract categories', and are then, as it were, projected onto different languages, whereas the specific grammatical and lexical means a language offers come with (sometimes subtle) differences in how such categories would or should be demarcated, and what interpretations are available. Verhagen therefore suggests to reframe

the question 'How does language X express STR-types A, B, C?' to 'What are the tools that language X makes available for the members of its community to manage the presentation of relationships between the mental states and feelings of different characters in a story, and the relationships of these to the narrator and the reader?', in order to refrain from presupposing types of phenomena independently definable of specific languages. As a construction is 'a pairing of a complex syntactic structure and a complex semantic structure' (Croft, 2001, pp. 203–204), this does not only raise the question whether or not types of conditionals are expressed using the same grammatical means (e.g., verb tense, modal marking, clause order) across languages, but also whether or not the same types of conditionals (and their demarcation) have evolved out of generalisations of conditionals used in usage events.

To acknowledge these difficulties, and to shed light on these matters, in what follows, I will offer a language-specific corpus-based account of conditionals in Dutch, which is not only my native language, but it has the advantage of having a vast body of literature available for investigating the linguistic features suggested in the literature discussed in the previous chapter.[9] It can also serve to test to which extent these features influence the implicatures of conditionals in languages other than English. Note that such an expectation is not far-fetched, and not at all at odds with the position defended by Croft (2001), as two Germanic languages will, of course, share characteristics, although they will not be identical. Instead of focusing on universality, I will focus on language-specificity. This means that the results of the second part of this dissertation, starting from the next sections, will primarily allow for conclusions about conditionals in Dutch, rather than about conditionals in general. This approach, and its benefits, can also be found in Verhagen's (2007) study, in which he shows that a comparative study between languages, English and Dutch in this case too, helps to gain a better understanding of the balance between regularities and idiosyncrasies in a language's grammar, and to get insight into the degree in which one 'complete grammatical system' overlaps with that of another language, and in which respects it differs. Notwithstanding, a language-specific study into linguistic features distilled from studies on another language appears, in this case, inevitable and while it may provide interesting insights as noted above, it also comes with the aforementioned risks. I will reflect upon these issues in further detail in chapter 7.

4.3.3 A bottom-up approach to conditional constructions

The initial approach to answering the research questions in this dissertation was to annotate both the types of conditionals distinguished in the accounts discussed in chapter 3, and the grammatical features inventoried. One could then determine to what extent the grammatical features are predictive of the

[9]This literature is mostly concerned with those features irrespective of their use in conditionals, but it will be discussed systematically in relation to each linguistic feature within conditionals in the next chapter.

implicatures of unassertiveness and connectedness using a so-called 'supervised machine-learning' approach. I will discuss this approach and an alternative approach in more detail in section 4.3.4 and in chapter 6, but in this section, I will provide four main arguments against a supervised approach for answering the aforementioned research questions.

First, as we saw in chapter 3, there are numerous accounts of conditionals, and in each of those accounts, different types are distinguished based on different criteria or different theoretical positions. Using the types distinguished for annotation would enable assessing which of the classifications is most likely to provide insightful groups of conditionals based on the distribution on their lower-level features. It would, however, to some degree, also assume these types and prohibit discovery of types that are not present in existing accounts. Connected to the argument for a language-specific corpus study above, annotating types of conditionals as discussed in accounts based on English furthermore assumes universal, or at least non-language specific types to exist. It is, however, not clear whether such an assumption is warranted. On the one hand, we do not know *a priori* whether types of English conditionals would have direct counterparts in Dutch. We also would have to assume that they show the same boundaries between types. While these are important questions, studies such as Renmans and van Belle (2003) and Reuneker (2017b) identified the types proposed by Dancygier and Sweetser (2005) in Dutch corpus data, and Verbrugge et al. (2007) provide experimental evidence for these types.[10] Even if we accept that the same types exist and are demarcated identically throughout languages, we would need to investigate to what extent their meaning can be attributed to the same grammatical means. On the other hand, as discussed briefly above, Croft's (2001) position strongly suggest negative answers to these questions, and although Verhagen's (forthcoming) conclusions are based on a study on perspectivisation, his conclusion that this phenomenon 'must not be framed in terms of relatively abstract categories of speech and thought representation, but in terms of interactions between the specific grammatical and lexical tools available in a specific language on the one hand, and the universal method of iconically depicting speech acts on the other' warrants caution for projecting language-independent categories onto language-specific conditionals too.

Second, if we were to annotate the types of conditionals discussed in the previous chapter, we would have to choose which classifications to use, because manual annotation is time-consuming and the number of accounts discussed in chapter 3 is large. Although a number of classifications show similarities, and all accounts discussed provide analyses of the same phenomenon, there are important differences between them. Although one could, for instance, choose

[10]See also the ample studies on coherence relations expressed by causal connectives, such as Scholman, Evers-Vermeul and Sanders (2016), and especially Sanders and Stukker (2012) for a cross-linguistic perspective on how causal connectives express relations in Sweetser's (1990) domains.

to use a selection of classifications that have been most cited and influential in the field, this would still amount to a biased choice, possibly discarding the most useful classifications.

Third, and what was admittedly the first reason the initial supervised approach was abandoned, concerns the reliability of annotation, as discussed extensively in the previous section. In accordance with the literature discussed, an experimental study by (Reuneker, 2017a; see previous section) showed that trained raters were able to reliably annotate modified textbook examples of implicatures of connectedness based on the accounts of Quirk et al. (1985), Athanasiadou and Dirven (1997a), and Dancygier and Sweetser (2005) (see section 3.3). However, the essential finding was that the same annotators were not able to reliably annotate the types of conditionals in actual language data. It is important to reiterate here that accuracy scores were used to select only competent annotators, and that these annotators turned out to be unable to classify corpus data reliably, with agreement scores ranging from 0.32 to 0.53. These coefficients indicate that annotations are not replicable between annotators, which means that they cannot be used as reliable data for further steps in the analysis. The results of the study suggest that generalisation, replication and meta-analysis may be compromised by the reliability of the classifications as applied to real language data. This is in line with what Carter-Thomas and Rowley-Jolivet (2008) suggest, namely that classifications of conditionals may be too idealised and detached from actual language use, possibly because the selected examples are not representative of all corpus data.

Fourth, by using a 'bottom-up' approach to classifying conditionals, i.e., by not using existing classifications of types of conditionals as labels, we can let the data speak for themselves. Although this may introduce the risk of presenting yet another classification of conditionals, as one cannot prevent identifying or discovering new types, and drawing boundaries differently (cf. Verhagen, forthcoming), the overview of types of conditionals and their grammatical features in chapter 3 minimises this risk, and maximises relating findings to existing types and features in the data.

The arguments above led to the decision not to annotate the higher-level types of conditionals, and rather to annotate the lower-level grammatical features inventoried. While many of these features, such as clause order or verb tense, are more explicitly marked and less interpretative than implicatures of unassertiveness and connectedness, other features, such as modality and aspect, are known to be more liable to ambiguity. Therefore, we will discuss the reliability of annotations of these features in the next section. First, however, we will flesh out the decision for a so-called *unsupervised* approach to data analysis in the next section.

4.3.4 Classification and clustering

As we saw in section 4.2, the annotation of the implicatures we are interested in, i.e., the types of conditionals discussed in chapter 3, proved problematic. Not only in terms of low reliability scores, but also in terms of biases introduced by selecting classifications to be used for annotation. This poses a problem for further analysis, as we cannot straightforwardly test which (combinations of) features are predictors of certain implicatures, which is what the research questions steer towards. In more methodological terms it means that we cannot, as is common in the field of machine learning, apply a classification algorithm to the features to be discussed in chapter 5, and see how well sets of features are indicative of types of conditionals, in effect testing the classifications discussed. These problems can and will be addressed, and in this section, I will briefly introduce the type of analysis that will be used to do so.

In the field of machine learning, there is particular interest in so-called *extensional classifications*. A large number of algorithms exists which take a set of features collected from observations or annotations (i.e., *multivariate analyses*), and consequently try to determine underlying classes of objects, which, in this study, would amount to types of conditionals. The two main approaches distinguished in the computational literature are *supervised* and *unsupervised learning*. The term 'classification' usually refers to what is called *supervised machine learning*. In this type of machine learning, the correct target labels (classes, types) for objects are known *a priori* for at least a number of observations (see e.g., Libbrecht & Noble, 2015). In contrast, unsupervised algorithms deal with data that lack such labels (see Berry, Mohamed & Yap, 2019, chapter 1). In other words, such algorithms involve pattern recognition without a target label, meaning that an algorithm is implemented to identify clusters of features inherent in the data, without any preconception of the nature of these clusters beyond the features that are used as input. So, whereas in supervised machine learning an algorithm tries to predict the correct label for an observation based on the distribution of features, trying to reach maximum accuracy, in unsupervised machine learning, no such target labels are available, which means that an algorithm has no clear, external labels to compare its results to.

Given the problematic reliability of annotations of types of conditionals in corpus data, no target labels are available for this study. This means that a supervised strategy is beyond reach, and the unsupervised technique of *cluster analysis* will be used and discussed in detail in chapter 6. Although this may seem a negative conclusion, the arguments in favour taking an unsupervised machine-learning approach are threefold, and relate directly to the arguments against classifying types of conditionals provided in the previous section. The first and most prominent argument, related to the third argument in section 4.3.3, is that it turned out that even trained annotators were not able to reliably classify conditionals in real corpus data. This means that there is no 'gold standard' (see e.g., Wiebe, Bruce & O'Hara, 1999) against which results

can be compared. The second argument, relating to the second argument in section 4.3.3, is that, even if reliable classification were possible, it is a non-trivial choice which classification or classifications should be used to provide the labels for the target attributes. This would introduce a theoretical bias, which unsupervised machine-learning does not suffer from, as there are no *a priori* class assignments. The results can be used to test to which extent the features suggested in the literature are indeed related to different implicatures of unassertiveness and connectedness. This is, in a sense, truly 'bottom-up', as unsupervised algorithms are forced to utilise the full potential of the data to find underlying structures (see also McEnery & Hardie, 2012, and chapter 6 on the corpus-based and corpus-driven distinction). The downside of this is that we are interested in specific implicatures, which is unknown to any algorithm to be implemented. However, given a constructional and pragmatic perspective, it may, as discussed in section 4.3.2, be expected that grammatical features give rise to these implicatures, but will not fully determine them. The unsupervised approach I argue for here provides a critical assessment of the relations between grammatical features and types of conditionals. The third argument, related to the first argument in section 4.3.3, is that it is an assumption that the types of conditionals distinguished in accounts based on English conditionals also exist in Dutch conditionals. Although Dutch and English are related languages, an unsupervised approach to conditionals in Dutch does not make this assumption, apart from the obvious and necessary selection of grammatical features used as input.

In conclusion, the above should be read as argumentation for and an introduction to the final, bottom-up analysis presented in chapter 6. The reason I discuss the approach here in this section is that it entails consequences for the corpus setup discussed next, and, in consequence, for the detailed discussion of the individual features in the next chapter. After all, the grammatical features of conditionals form the input for the cluster analysis presented in chapter 6, with which we will try to measure the extent to which grammatical features of conditionals in Dutch form clues for implicatures of unassertiveness and connectedness, i.e., to which extent such an analysis provides a foundation for a meaningful, data-driven groupings of conditionals.

4.3.5 Conclusion

In this section, we discussed the arguments for a corpus-based approach to conditionals in light of the research questions presented at the end of chapter 2. Based on, among other factors, the problematic reliability scores of annotation types of conditionals, an unsupervised, bottom-up approach of cluster analysis was chosen in this section as the most promising method of uncovering relations between the grammatical form and meaning of conditionals. The input for such an analysis ideally consists of high-quality annotations of corpus data, which we will turn to next by addressing the corpus setup in the next section, and the annotation of corpus data in section 4.5.

4.4 Corpus setup

4.4.1 Introduction

In this section, I discuss the setup of the corpus used in this study, with special attention to the representativeness of the language data used, and the sampling strategies used to arrive at a balanced corpus that allows for specificity as well as generalising of conclusions.

The aim of this section is to provide a clear picture of the data used in the remainder of this study. In section 4.4.2 I will discuss which population the corpus study targets to describe, for which we will look at mode and register in section 4.4.3. Section 4.4.4 is devoted to the identification of conditionals in Dutch, which is less straightforward than in English. In 4.4.5 I will present the final sampling frame, and in 4.4.6, finally, I will offer a brief conclusion before moving on to the annotation of the corpus materials.

4.4.2 Population and representativeness

In this section, I discuss what population the corpus study aims to describe, or, in other terms, what the actual object of study is for answering the research questions. As existing corpus studies show (see Ford & Thompson, 1986; Ferguson, 2001; Carter-Thomas & Rowley-Jolivet, 2008), the use of conditionals differs significantly between modes (spoken, written), genres (e.g., newspaper press, discussion fora, academic texts) and registers (formal, informal). When one strives for maximum representativeness in the sampling frame, it is important, therefore, to ensure that findings can be taken to represent characteristics of the population. Any study, therefore, needs to address what is taken to be the population of interest, i.e., the full range of phenomena of interest, such as 'spoken language' or even 'language' (see e.g., Buchstaller & Khattab, 2014, p. 74).

In most (corpus) linguistic studies, it is not possible to investigate the whole population directly, which means that appropriate samples must be constructed. Before the samples can be constructed, however, the population itself must be defined. When a researcher is interested 'only' in academic writing, child-directed speech or doctor-patient interactions, this determines the population to be sampled. Examples of corpus studies in which this is possible are those focusing on, for instance, the use of certain linguistic features in the complete works of one author, such as the corpus-stylistic work on Dickens's novels by Mahlberg and Smith (2012). The current study, however, does not limit the population to a specific author, genre or register. The sampling frame should therefore be properly defined with respect to a target population (cf. Atkins, Clear & Ostler, 1992). A target population is a complete set of observations that share at least one characteristic (see e.g., Banerjee & Chaudhury, 2010). The target population of this study is defined as all conditional *als*-sentences in Dutch, as spoken and written in the Netherlands. This definition excludes

several regions in which Dutch is used, such as Belgium, Suriname, Aruba, Curaçao and Sint Maarten. The reason for this exclusion is to limit the influence of regional variation.

As may be expected from this definition, it is not possible to access the target population directly. The accessible population (Bracht & Glass, 1968) (also called 'study population', cf. Banerjee and Chaudhury, 2010) is the population that is available for sampling, and is defined as follows: all sentences in which the conditional conjunction *als* is used and that are available in existing corpora of spoken and written Dutch. Notice that conditionals in this definition are limited to those introduced by the conjunction *als*, which was decided in order to have the most direct link to the classifications and features discussed in the previous chapter, and to have a baseline of the default conditional in Dutch to which, in future research, more specific conditional constructions can be compared. The accessible population will be used as reference for the sampling frame, meaning that the conclusions based on the data used in this study are intended to be indicative for Dutch in the Netherlands (the target population) through the accessible population by means of the samples discussed below.

With respect to the studies mentioned above that suggest differences in use of conditionals between modes and genres, it is important to consider the sample representativeness of this study. A representative sample is defined as a sample that has the same distribution of features as the population it was taken from. The notion of 'representativeness' is a relative notion, i.e., a sample is representative for a particular population, selected by the researcher (see e.g., Sankoff, 1989). As McEnery and Hardie (2012, p. 10) point out, however, representativeness is an ideal that is 'rarely, if ever' attained. Nevertheless, Leech (2007, p. 143) argues that while the goal of representativeness may not be achieved in full, we should not abandon pursuing it: 'we should aim at a gradual approximation to these goals, as crucial desiderata of corpus design'. In the design of the corpus for this study, I aim to be maximally explicit about the sampling frame and its relation to the population, while acknowledging that perfect representativeness is out of reach, because, among other factors, it is hard to determine and quantify the level of representativeness (cf. Biber, 1993).

As I argued above, I do not intend to limit the object of study to a particular context of use, because any conclusion would then be limited to this context. A clear example of this approach is Ferguson (2001), who provides a detailed analysis of the uses of conditionals in doctor-patient interactions, which provides insights into the use of conditionals in specific contexts. The conclusions cannot easily be generalised to other contexts, however. To be clear, this sample may still be representative, but only for a narrowly defined target population. Rather, I will strive for 'generalisability' instead of specificity by constructing a maximally representative corpus relative to a widely defined target population. To make sure the samples are balanced, the sampling procedure I adopt here is 'random stratified sampling' (see e.g., Rice, 2010, p. 240). The reason for this is that, as the studies cited above have shown, mode, genre and register

are variables that influence the use of conditionals. As these variables are not evenly distributed in the corpora (see below) and, of course, nor in language in general, I have taken care to represent these 'stratifying variables' evenly in the sampling frame.

4.4.3 Balance, mode and register

The corpus study presented here hosts two main strata: spoken and written language. The randomised selection of conditionals from spoken Dutch is comparable in size to the selection of conditionals from the written corpus. Next to the reason provided above, an added benefit is that in most corpus studies, spoken language has, at its most, a subordinate role (see e.g., Gabrielatos, 2010; Reuneker, 2016; for exceptions, see e.g., Ford & Thompson, 1986; Athanasiadou & Dirven, 1995). This has to do with the availability of spoken natural-language data, because written publications – from digital to digitised – are much less time consuming to use as material for a corpus. An added benefit of a stratified approach is that it enables the direct comparison between samples. If a feature is more frequent in the spoken sample than in the written sample, this may be interpreted as evidence for a difference between those two sub-populations. This would not be equally possible with fully random sampling, given the considerable size differences between the two source corpora, which we will discuss next.

The two corpora I used for data collection are the most recent corpus of spoken Dutch, the 'Corpus Gesproken Nederlands' or CGN (Oostdijk, 2000), and the most recent corpus of written Dutch, the SoNaR corpus (Oostdijk et al., 2013). The CGN hosts almost 9 million words (van Eerten, 2007), whereas the SoNaR corpus hosts over 500 million words (Oostdijk et al., 2013, p. 222). Not all sections of these corpora were used, as I will discuss below.[11] Care was taken to include in the results both the linguistic context of the sentences (i.e., the preceding and following sentence), as well as the necessary metadata.[12]

Biber (1995) argues that a distinction between written and spoken language may be too broad, because these modes may be similar in some respects, but very different in others. Some differences between texts may be connected more to other dimensions than mode itself. The distinction spoken-written is upheld here, however, because previous discourse-oriented studies (Carter-Thomas &

[11] Note that both corpora feature language use from both The Netherlands and the Dutch-speaking part of Belgium (Flanders). While in the CGN approximately 66 percent of the material is recorded in The Netherlands and 33 percent in Belgium (see Oostdijk, 2000, p. 280), for SoNaR these figures are approximately mirrored (see Oostdijk et al., 2013, p. 244). As I included only data from The Netherlands in this corpus study, these differences are not relevant, but it does mean that not the full corpus sizes should be considered as the accessible population, but the proportions just reported.

[12] The examples from these corpora are presented in this dissertation together with a reference to their origins within the respective corpora. Labels starting with (lowercase) *fn*, as in *fn000149*, indicate an example comes from the CGN, whereas labels starting with capital letters, as in *WR-P-E-A-0005795081* indicate an example comes from the SoNaR corpus.

Rowley-Jolivet, 2008; Ford, 1997; Ford & Thompson, 1986) have shown that, in the case of conditionals, the spoken-written language dimension is relevant. Phenomena like hedging, insubordination and clause order seem related to mode more than to other parameters. Because Biber and Conrad (2009, p. 88) argue that 'the language of conversation is highly distinctive compared to the language of books', I chose to define further strata within both modes. Spoken conversation, for instance, may differ more from spoken language in formal debates than from informal written texts on discussion boards (see also the notion of 'hierarchical sampling strata' in Biber, 1993, p. 244).

As the sub-populations vary in size, the main set-up of the corpus study is multistage sampling in a stratified design. This means that the sub-populations are divided into homogeneous subgroups or *strata*, which will all be sampled independently. As Biber (1993, p. 244) argues, 'stratified samples are almost always more representative than non-stratified samples', because the strata can represent the proportions desired, instead of relying on random sampling. Note that I explicitly choose a stratifying approach here to be able to investigate differences between sub-populations that have been shown to differ with respect to the use of conditionals (see above). The downside is that the collective samples cannot be taken to directly represent the population together, as we do not know their exact distribution in language. The upside is that there is no risk of having a sub-population in random sampling dominate the results, 'just' because it forms a larger part of the corpus. This is a real risk, as in most corpora, written texts, such as newspaper texts and, more recently, discussion list texts, make up for the majority of the corpus. In strata, the within-group variance is typically smaller than the between-group variance, representing both the sub-populations and the whole population better. To ensure that the corpus sections I selected vary systematically within the spoken and written modes, two further dimensions distinguished by Biber (1988) are used.

The first parameter on which strata are defined is the dimension 'involved vs. informational production' (Biber, 1995, pp. 141–151). Involved language use is highly interactional and features high frequencies of private verbs (*know*, *think*), *that*-deletion, contractions, present tense verbs and second person pronouns, whereas its mirror-image is informational language use, featuring a higher type-token ratio, greater word length, and high frequencies of nouns and prepositions. With respect to the arguments concerning the choice of a specific language, it may be argued that these features can be highly language-specific. In general, however, it may be expected that these dimensions may influence language use in other languages.[13] The second parameter used is Biber's third dimension, 'Situation dependent vs. elaborated'. Situation-dependent language

[13]For instance, recent findings by van Beveren, Colleman and de Sutter (2018) show how register affects the use of the optional prepositional complementiser *om* 'to' in Dutch infinitival complements, as in (a) below.

(a) Ik beloof (*om*) op tijd te komen.
I promise *to* be on time. (van Beveren, Colleman & de Sutter, 2018)

scores high on time adverbials, place adverbials, and other adverbs, while elaborated language features more *wh*-relative clauses on object and subject positions, phrasal coordination and nominalisations (Biber, 1995, pp. 155–159). The reason for choosing these two parameters is that results of the corpus studies previously mentioned, albeit not in such specific terms as textual dimensions, indicate differences in usage of conditionals between genres such as academic and advertorial writing.

Because a large-scale multidimensional analysis of the corpora is outside the scope of this study, Biber's dimensions were used to identify the most appropriate counterparts of Biber's samples in the corpora used as data source. This comparison is only used as a proxy to verify the identification of corpus segments comparable to Biber's registers. For instance, while Biber's register of 'face-to-face conversations' can be matched directly with 'spontaneous face-to-face conversations' in the spoken corpus, 'official documents' cannot, as the written corpus hosts categories that are related, but not identical to Biber's registers, such as 'policy documents' and 'proceedings'. For each of the dimensions, the most similar available sections were chosen from the corpora. For instance, on the dimension 'involved vs. informational production' the register with the highest mean score on features that add up to 'involvedness' is 'face-to-face conversations' (Biber, 1995, p. 117). The register with the lowest mean score on that dimension is 'academic prose' (Biber, 1995, pp. 118–146). For this dimension, thus, 'involved language use' includes genres such as 'spontaneous face-to-face conversations' for the spoken mode and 'discussion lists' for the written mode, while 'informational language use' includes 'news reports' for the spoken mode and 'manuals' for the written mode. In section 4.4.5 the final sampling frame, including the original corpus sections, aree presented, but first, in the next section, the distinction between conditional and non-conditional *als* will be elaborated, as it is needed for the identification of conditionals from the selected corpus components.

4.4.4 Identification of conditional *als*-sentences

For each of the samples, all sentences were randomly ordered and I identified which sentences featured *als* 'if' as a conditional conjunction, until the desired number of sentences for each of the samples was found (see below). This was done because the conjunction *als* 'if' can be used in several ways in Dutch. Furthermore, as Pollmann (1975, p. 187) argues, 'many *als*-sentences have more than one interpretation'. Because of this, the identification of *als* 'if' as a *conditional* conjunction is less straight-forward then it is for English *if* (see e.g., Declerck & Reed, 2001, p. 9; Gabrielatos, 2010, p. 45). The procedure of identifying conditional use of *als* 'if' is therefore discussed in detail in this section.

Each of the uses of *als* 'if' as a conjunction distinguished by de Rooy (1965) will be discussed, because his account clarifies how conditional sentences can be distinguished from non-conditional sentences in which *als* 'if' occurs as a

conjunction. Next to its comparative use, de Rooy distinguishes between its use as a conjunction of manner, a conjunction of qualification (or 'state of being'), a temporal conjunction, and, finally, a conditional conjunction. [14]

Before discussing these different uses, a remark on reliability is in order. In what follows, I discuss the relevant literature that was used to identify conditional use of the conjunction *als* 'if'. This does not mean, however, as we will see below, that no ambiguous cases remained, or no errors could have been made. Although no study of the reliability of this selection procedure was performed, as was done in the experiment reported on before, and as is done for the annotation of features presented in the next chapter, during the annotation of those features, the second annotator was instructed to comment on uses of *als* 'if' that, according to him or her, did not qualify as conditional use. Although this does not amount to a full assessment of inter-annotator agreement for the identification of conditional *als* 'if', together with the explicit discussion of criteria for the conditional use of *als* 'if', it believe the approach was sufficient. Nevertheless, extending the evaluation of reliability to the identification of conditional *als* 'if' is suggested here as an improvement for future research.

The first use is *als* 'if' as a comparative conjunction, as in (9), comparing a noun phrase to a noun phrase, and (10), comparing an adjectival phrase to a noun phrase.

(9) Die man heeft een leven *als* een prins. (de Rooy, 1965, p. 144)
That man has a life like *a prince.*

(10) Onze metselaar is zo dik *als* een pad. (de Rooy, 1965, p. 144)
Our bricklayer is as thick as *a toad.*

Als 'if' in (11) and (12) below are used as comparatives as well, although the former is infrequent or regional, as the regular conjunction in this use is not *als* 'if' but *zoals* 'like' (see a.o. Overdiep, 1937, p. 590; Haeseryn et al., 1997, pp. 567–570). *Als* 'if' in (12) is an example of 'incorrect usage' according to prescriptivists (see e.g., Charivarius, 1943, p. 35; see, for descriptive accounts, Paardekooper, 1950; Paardekooper, 1970; Postma, 2006; Stroop, 2011; Hubers and de Hoop, 2013), because in unequal comparisons, *dan* 'than' is prescribed.

(11) Het is *als* je zegt. (de Rooy, 1965, p. 144)
It is as/like *you say it is.*

(12) Hij is groter *als* zijn broer. (de Rooy, 1965, p. 144)
He is larger as *his brother.*

[14] I will ignore the use of *als* 'if' as explanatory conjunction, as in *de kloosterlingen, die, als van Frankische afkomst, meest ongeleerd waren...* 'the monks, who, as of Frankish descent, were most unlearned ...', which, was not found in the dialects studied by de Rooy and, with respect to standard Dutch, 'will probably be limited to special styles' (de Rooy, 1965, p. 37). Furthermore, *als* 'if' used as an expletive (redundant) conjunction, as in *hij zou* als *morgen komen* 'he would *{as/if}* come tomorrow', is ignored as well, as it is considered regional and archaic (de Rooy, 1965, pp. 65–67).

In (13), *als* 'if' functions as conjunction of manner, i.e., 'the air was as it would be if it were wiped clean', and in (14) as a conjunction of qualification (*hoedanigheid* 'state of being'), i.e., 'He rules in his function of king'.

(13) De lucht was op eenmaal *als* schoon geveegd. (de Rooy, 1965, p. 33)
The air was at once as *wiped clean.*

(14) Hij regeert *als* koning. (de Rooy, 1965, p. 36)
He reigns as *king.*

From (15) to (18) the examples become more relevant to the current discussion, as in all these cases, the conjunction *als* 'if' introduces not a phrase, as in the examples above, but a complete clause (subject and predicate), which, by means of the conjunction, is subordinated to the main clause of the complex sentence. This does not mean, however, that all of these examples are conditional sentences.

(15) *Als* ik gisteravond thuiskwam, waren de anderen al naar bed. (de Rooy, 1965, p. 144)
If [when] *I came home last night, the others were already in bed.*

(16) *Als* de kippen een sperwer zien, zijn ze bang. (de Rooy, 1965, p. 144)
{If/When} *the chickens see a sparrow hawk, they are scared.*

(17) *Als* ik jou was, zou ik het doen. (de Rooy, 1965, p. 144)
If *I were you, I would do it.*

(18) *Als* hij het maar deed! (de Rooy, 1965, p. 144)
If only *he did/would do it!*

In example (15), *als* 'if' can only be read as introducing a temporal relation between coming home in the subordinate clause and the others having gone to bed in the main clause. Such a purely temporal use is described as 'non-standard Dutch' by de Rooy (1965, p. 143; see also Pollmann, 1975, pp. 188–189). Overdiep (1937, pp. 588–589) mentions this use and connects it to the use of the past tense and the historical present, as in (19) and (20) respectively.

(19) We waren in dien tijd niet verwend! Als om acht uur de postwagen langs reed en de horen door de straten schalde, dan kregen we allen een schok van blijde verrassing. (Overdiep, 1937, p. 588)
We were not spoiled at that time! {If/When} *the mail wagon drove past at eight o'clock and the horn blew through the streets, we all got a shock of happy surprise.*

(20) Verbijsterd stáán (= bleven staan) zij vervolgens in hun loop als achter hen Ballochi's wakk're troep aanstormt [...]. (Overdiep, 1937, pp. 588–589)
Standing stunned in their course as Ballochi's awake troop storms behind them.

Typically, *als* 'if' such as in (19) are used to refer to recurring events ('everytime the mail wagon drove past...'). This can also be seen in (21), in which *meestal* 'usually' highlights the recurring nature of the event.

(21) Zes treffers vielen in de mislukte kwalificatie voor het WK, allemaal tegen de kleinere landen Estland, Andorra en Cyprus en meestal *als* het duel lang en breed was beslist. (WR-P-P-G-newspapers-138000)
Six goals were made in the failed qualification for the World Cup, all against the smaller countries of Estonia, Andorra and Cyprus, and usually if *[when] the game was decided already.*

In case of reference to a singular event, *toen* 'when' is used, and *als* 'if' as in (19) is deemed 'irregular' by Overdiep (1937).[15] In Belgian-Dutch, however, the temporal use of *als* 'if' as in (15) and (19) is common (Haeseryn et al., 1997, pp. 553–554). As the corpus only contains Dutch from the Netherlands, only a few instances of this use were found and they were discarded from the samples. The historical present, as in (22), is found mostly in narrative contexts in Dutch (for a recent account, see Sanders & van Krieken, 2019). These backshifted contexts are easily recognisable, and were subsequently excluded from the samples.

(22) 'Dat is uiteindelijk toch het probleem met een eenpartijstaat', zegt Sie met dat zangerige Indonesische accent van hem, *als* we dat [*sic*] toch aan tafel kunnen schuiven. (WR-P-P-G-newspapers-98000)
'Ultimately, that is the problem with a one-party state,' says Sie with his vocal Indonesian accent, if *[when] at last we can gather round the table.*

Moving on to the example in (16), this use of *als* 'if' is termed 'temporal-hypothetical' by de Rooy (1965, p. 143), and Overdiep (1937, p. 588) too considers this temporal use, not conditional use of the conjunction. However, as became clear from the previous chapters (see the accounts by Sonnenschein in section 3.2.4, and by Athanasiadou and Dirven in 3.3.9 specifically), I do consider this usage conditional here. In terms of Athanasiadou and Dirven (1996), this would even be a prototypical example of a *course-of-event conditional*. The difficulty here is that it is possible, or even likely, that a temporal relation between antecedent and consequent, over time, develops into a conditional relation through regularity, i.e., there may be a gradual transition from a purely temporal relation (p before q), to a regular temporal relationship (p often before q), and finally to a more systematic relation, such as a rule or law (whenever p, q). For the latter relation, p may finally be construed as the cause of or condition for q (on the notions of regularity and causality, see e.g., Lewis, 1973a; Schulz, 2011, pp. 14–15). This hints towards a continuum rather than a strict temporal-conditional dichotomy in Dutch conditionals, and given the accounts by Athanasiadou and Dirven and others mentioned above, I will

[15] Original text: 'Ongewoon is in Noord-Nederl. de functie van aanduiding eener enkele, momentane, handeling (gewoon is hier: *toen* [...])' (Overdiep, 1937, p. 588).

exclude only very clear temporal uses of *als* 'if' as non-conditional. Although there are important differences between Dutch conditional *als* 'if' and temporal *wanneer* 'when' on the one hand, and English conditional *if* and temporal *when* on the other hand, it is important to note, as Dancygier and Sweetser (2000, p. 112) do, that both the similarities and differences between conditional and temporal conjunctions should be analysed not by 'focusing only on the conjunctions themselves, but by describing the range of constructions they participate in'. Here, we focus on conditionals expressed using *als* 'if', but as the aforementioned temporal-conditional continuum may be associated with the further grammatical features of *als*-constructions, such as the extent of modalisation of the consequent, a comparative analysis of conditional *als* 'if' and temporal *wanneer* 'when', which unfortunately falls outside the scope of this dissertation, may shed light on this matter. Such an analysis is suggested for future research and discussed in more detail in chapter 7. The type of conditional found to be most central in many accounts is found in (17), which de Rooy calls 'hypothetical', and it is the only type he describes within the category of *als* 'if' as conditional conjunction (de Rooy, 1965, p. 56). Finally, in (18), we see the optative use of *als* 'if' in an insubordinate clause. (For an account of Dutch insubordinate conditionals, see Boogaart and Verheij, 2013; and for an account of insubordinate conditionals in Germanic languages, see D'Hertefelt, 2015, Chapter 2.)

The discussion above makes clear that simply isolating sentences with *als* 'if' and filtering out some known non-conditional uses, as can be done for English *if* (see Declerck & Reed, 2001, p. 9, and Gabrielatos, 2010, p. 45; see also section 2.2) will not suffice for Dutch. I hope to have shown here that manual inspection and selection of corpus data is necessary.[16] A welcome by-product of this strategy is that it forces the researcher to more clearly define beforehand what grammatical pattern is needed for *als* 'if' to receive a conditional reading. What distinguishes the conditional examples above from the non-conditional examples from a syntactic perspective, is that the sentences are complex (involving a subordinate and a main clause or, in case of insubordination, only an insubordinate clause), which is connected to the first criterion of the preliminary characteristics I presented in section 2.2, i.e., conditionals are 'bi-partite'. As that criterion needed to include conditionals expressed by other means than *als* 'if', for the selection of conditional *als*-sentences it can be sharpened here into the criterion of the sentence being 'bi-clausal'. The use of the conjunc-

[16] I would like to remark here that the search capabilities for the corpora used here have been extended during the duration of this project. For this project, I have indexed all texts and converted them into a format easily readable in Python (van Rossum and Drake, 2009; see Appendix F). Still, as far as the metadata, such as POS-tags, go, it is still not possible to separate conditional from non-conditional *als* 'if', which is, as mentioned, different for English conditionals, although formulating regular expressions to identify conditionals *if* involves its own challenges, such as excluding indirect interrogatives with *if* is quite tricky, 'as verb forms other than the bare form may well return conditionals', and manual cleaning continues to be an important and necessary step (C. Gabrielatos, personal communication, September 11, 2015).

tion *als* 'if' allows for distinguishing between conditional sentences and other bi-clausal sentences. As we have seen in the preceding chapters, *als* 'if' adds unassertiveness and connectedness to the expression of propositions p and q (i.e., the second and third characteristics). Furthermore, the subordinate clause, as is expected, has the finite verb in clause-final position.[17] Supplementing these characteristics with the discussion in this section allowed for distinguishing between conditional and temporal use of *als* 'if', although this, as discussed, remains an interpretative endeavour to a certain extent. On a side note, although all of the conditional examples in this section have sentence-initial subordinate clauses, this is not necessary. For instance, the conditional clause in (17) can easily be placed in sentence-final position and, somewhat less easily, in sentence-medial position, as can be seen in (23) and (24) below.

(23) Ik zou het doen, *als* ik jou was.
I would do it, if *I were you.*

(24) Ik zou het, *als* ik jou was, doen.
I would, if *I were you, do it.*

We will discuss variations of clause order in detail in section 5.2. Based on the discussion above, it becomes clear that we are interested primarily in the type in (16), (17) and, to some extent, (18).

One problem needs to be addressed before moving on to the actual sampling frame. Whereas in English, *if* cannot, or only in a very limited range, be used for purely temporal relations (Dancygier, 1998, p. 48; Declerck & Reed, 2001, pp. 31–5), in Dutch, it is customary to use *als* 'if' for non-conditional, purely temporal relations, as in (25) below.

(25) *{Als/Wanneer}* je morgen wakker wordt, krijg je een cadeau.
{#If/When} *you wake up tomorrow, you will get a present.*

In English, it would be mandatory to use *when* instead of *if*.[18] While this example is clear, this is not always the case. As Overdiep (1937, p. 589) argues, '[this type of] adverbial temporal clause introduced by *als* almost inevitably describes a future event; therefore the function of the adverbial temporal clause is hard to distinguish from that of the adverbial conditional clause'. In the historical dictionary of Dutch *WNT*, this ambiguity is observed too: 'Not always unambiguously separable from conditional *als*'.[19] The difference between *als* 'if' and *wanneer* 'when' as conditional and temporal conjunctions is, as remarked in a footnote, not pursued any further by van Belle (2003, p. 67), although he

[17] 'Clause-final position' is used here in the sense of Broekhuis and Corver (2016, pp. 1245–6), who argue that the term should not be taken to mean that the finite verb 'demarcate[s] the right boundary of the clause', as it can be followed by other constituents, such as prepositional phrases. Rather, it is taken to mean that the finite verb is 'in the right periphery of the clause'.

[18] Although this does not mean *when* is never used in English for the expression of a conditional. It is listed in the *Top 50 Grammar Mistakes* (Wallwork, 2018, pp. 41–43).

[19] Original text: 'Niet altijd ondubbelzinnig te scheiden van bet. 1.1.2 "als, wanneer"'.

does mention that the conditional use of *wanneer* 'when' is more formal than *als* 'if', and that *wanneer* 'when' cannot be used in counterfactuals, such as in his example reproduced in (26) below. Duin (2011), however, presents an attested counterexample, as adapted in (27).

(26) {*Als/?Wanneer/Indien*} de marsmannetjes ons overvallen, blijft niemand van ons in leven. (van Belle, 2003, p. 67)
{If/?When/In case} *the Martians are attacked, none of us will live.*

(27) Een verdiende zege voor DZC'09 die zelfs nog hoger had kunnen uitpakken *wanneer* ze het nog wat slimmer hadden uitgespeeld. (Duin, 2011, p. 25)
A deserved victory for DZC'09 that could have turned out even higher {if/?when} *they had played a little smarter.*

In fact, more counterexamples can be found.

(28) Het zou Wellink én Cornet hebben gesierd wanneer ze hadden ingezien dat een gentleman die zes maanden in zijn eigen levensonderhoud zou hebben voorzien. (WR-P-P-G-0000004603)
It would have made Wellink and Cornet look good {if/?when} *they had realised that a gentleman would have provided for himself for six months.*

This is in line with the argument in section 2.5.4, in which I argued, along the lines of Karttunen and Peters (1979b, pp. 5–6), Langacker (2008, p. 302), and Dancygier and Sweetser (2005, p. 76), that subjunctive conditionals are better described in terms of implicatures of epistemic distancing than in terms of (semantic or presuppositional) counterfactuality. While, as Duin (2011, p. 36) shows, *wanneer* 'when' may be used less often in counterfactual contexts, it is not wholly incompatible. Other factors, most notably considerations of style and formality, are of influence. This point will, for reasons of space, not be taken up further here. The detailed classification of conditional-temporal *als*-sentences by Pollmann (1975) makes clear in which contexts the ambiguity in question arises. For Pollmann, the example in (29) below is ambiguous, because it can be used to express either the speakers certainty about the guests coming tomorrow (i.e., the temporal, *when* interpretation), or to express the speakers uncertainty about the guests coming (i.e., the conditional, *if* interpretation).

(29) *Als* de logés morgen komen, vinden ze de kamers op orde. (Pollmann, 1975, p. 190)
{If/When} *the guests arrive tomorrow, they (will) find the rooms in order.*

Pollmann shows that this ambiguity is highly context-dependent and can shift by switching between definite and indefinite descriptions (i.e., 'If *guests* come tomorrow' and 'If *the guests* come tomorrow') and the place of time adverbials, as in the difference between (29) and (30) below.

(30) Als morgen de logés komen, vinden ze de kamers op orde. (Pollmann, 1975, p. 189)
{If/When} *tomorrow the guests arrive, they (will) find the rooms in order.*

In this example, Pollmann (1975, p. 190) argues, the speaker does not take into account the possibility that the guests might not come. As can be seen, these judgements are not clear-cut and in the samples used in this study too, a number of such ambiguous cases was found, such as (31).

(31) Het recht op toekenning of behoud van de persoonlijke garantietoeslag vervalt als de werknemer [...] met (vroeg)pensioen gaat. (WR-P-P-F-0000000014)
The right to grant or retain the personal guarantee allowance lapses if the employee [...] takes (early) retirement.

As this example concerns retiring, which is tied to a certain age, the antecedent can be interpreted as 'the moment in time the employee retires', but here, it seems that retiring is presented as a condition for loosing the right on an allowance. Although this specific example was treated as conditional, both interpretations can be argued for. As context was included in the samples, in most cases, the preceding texts were used to exclude certain interpretations. A final point related to this discussion is that, as we saw, the uses of *als* and *wanneer* do not coincide with the conditional-temporal distinction to the same extent as English *if* and *when*. Including purely temporal uses of *als* 'if' in the corpus, therefore, would amount to including a specific use of Dutch *als* not found (to the same extent) for English *if*, as it would be expressed by another conjunction (*when*).[20] Given that the literature used to identify types of conditionals and their suggested linguistic features was based on English *if* only, I chose here to limit the included uses of *als* 'if' to conditional uses.

In this section, I discussed which sentences containing the conjunction *als* 'if' are included as conditionals in the corpus. We have seen which criteria were used for the identification of conditional *als*-sentences. In the next section, I will present the final sampling frame, before continuing with the discussion of the annotation of data in section 4.5.

4.4.5 Final sampling frame

As *als* 'if' is the default conditional conjunction in Dutch, each sample contained a sufficient amount of occurrences. While sample size is important, it is often hard to calculate exactly the number of observations needed for a study to be considered representative and include an appropriate dispersion of the variables involved. The strategy followed here was to strive for a sample of 5.000 *als*-conditionals and to check for dispersion of individual features. This number was chosen by practical argument mostly, as it was large enough for the quantitative analyses presented in the next chapters, while manual annotation was still feasible. This is not to say that there was no more systematic or principled way of evaluating this sample size. For this, as we will see in the next chapter, I have

[20] As we will see below, not only can *als* 'if' be used to express temporal relations, but *wanneer* 'when' can also be used to express conditional relations.

performed an initial annotation of a random sub-sample of 500 conditionals to check for dispersion of features. The feature value with the lowest frequency was second-person plural subjects (see section 5.7), which, in this sub-sample, occurred only 2 times in all antecedents and 2 times in all consequents. Crucially, although these numbers are, of course, low, it did indicate that the final sample of 5.000 conditionals would not be void of any non-occurring individual feature values. Before the final sampling frame is presented, including the intended and realised frequencies within each sample, a remark about the sample 'discussion list' is in order.

Whereas the full sample of discussion list data in the SoNaR corpus is gathered from a number of sources, almost all data from the Netherlands appear to be gathered from the discussion list section of the website *Ouders Online* 'Parents Online', which is mainly used by (soon-to-be) parents. While these data are valuable, this poses problems for the representativity of the respective sample. For instance, almost all discussions in the discussion list data were between women in a narrow age range and revolved around the theme of having babies, raising children and relational problems with partners and parents-in-law. To solve this problem, the administrators of several large Dutch discussion lists were contacted to ask for a sample of their data. Given the time in which this was done – around the same time the General Data Protection Regulation was heavily covered in the news – administrators were reluctant to provide even anonymised, sampled data.[21] The technology-oriented discussion list *Tweakers*, however, was willing to supply data. While this is not ideal (i.e., more sources would have been preferred), most of the discussion list data in the sampling frame below now comes from not one, but two sources. The upside is that the demographics and topics of both discussion lists differ significantly. A second addition concerned the sources for the formal written texts in the SoNaR corpus, which are largely newspaper articles, newsletters and press releases, whereas legal texts and policy documents are limited. Therefore, I have added texts from five academic journals to the sampling frame, to reach the required number of formal texts outside newspaper texts.[22]

With the remarks above in order, the sampling frame is presented in Table 4.5 below.

[21] See https://gdpr.eu. I have consulted a Leiden University lawyer (personal communication) on this matter, who ensured anonymous, non-traceable data would not introduce any legal issues.

[22] The texts were extracted from the Dutch academic journals *Nederlands Tijdschrijft voor Geneeskunde* 'Dutch Journal of Medicine', *Tijdschrift voor Geschiedenis* 'Journal of History', *Algemeen Nederlands Tijdschrift voor Wijsbegeerte* 'General Dutch Journal for Philosophy', *Nederlands Tijdschrift voor Handelsrecht* 'Dutch Journal of Commercial Law', and *Tijdschrift voor Criminologie* 'Journal of Criminology'. Linguistics journals were excluded to prevent inclusion of references to linguistic phenomena and metalinguistic terminology.

Table 4.5:
Final sampling frame

Mode (2500)	*n*	Register (1250)	*n*	Genre (625)	*n*
Written	2462	Formal, informational	1240	Newspaper, newsletter, press release	690
				Legal, policy, academic journal	550
		Informal, involved	1222	Discussion list	605
				Chat, SMS	617
Spoken	2406	Formal, informational	1186	Broadcast news	600
				Political discussions	586
		Informal, involved	1220	Spontaneous conversations	599
				Telephone conversations	621
Total					4868

Note. Targeted frequencies per dimension are represented between parentheses.

As we can see here, due to data selection, not all intended frequencies were achieved in full, but a total size of 4868 was deemed close enough to the intended 5000 conditionals.

4.4.6 Conclusion

In this section, we reviewed the considerations that led to the design of the current corpus. I have discussed the necessary steps to assure representativeness by means of a well-balanced corpus. As became clear in chapter 2, there is no consensus on a clear definition of conditionals in natural language, and atop that, in Dutch, it is not always possible to unambiguously distinguish between conditional and non-conditional use of the conjunction *als* 'if', especially in relation to the temporal use of *als* 'if'. We therefore reviewed the identification of conditional *als* 'if' in the corpus, and finally, we discussed the sampling frame. The next section discusses the annotation of the features identified in the previous chapter, which we will turn to now.

4.5 Corpus annotation

4.5.1 Introduction

In the next chapter, the distributions of features identified in the previous chapter will be investigated. Although automatic annotation of grammatical features is preferred (see e.g., Levshina & Degand, 2017), if it can be done reliably, most of the relevant features carefully identified in the previous chapter were not available for such pre-processing. Features were, for the largest part, manually annotated. Therefore, I deem it necessary to elaborate the annotation process and, with an eye on the experiment presented in section 4.2, to critically assess the reliability of the annotation. In this section, I discuss the notion of agreement briefly, and especially the measures to ensure maximisation of reliability.

4.5.2 Reliability measures

As we saw in section 4.2, the application of classifications to natural language conditionals did not produce high reliability scores, (partly) due to the fact that the classifications tested represent coherence relations that are not often explicitly marked in conditionals. This problem of linguistic underspecification extends to lower-level features, for which ambiguities may arise as well, such as in the case of modal verbs. A clear example is provided by Boogaart and Reuneker (2017). The modal verb *must* can be used to express deontic modality, as in (32) below, or to express epistemic modality, as in (33).

(32) He must be home by 6, so he should really go now. (Boogaart & Reuneker, 2017, p. 199)

(33) He must be home since the lights are on. (Boogaart & Reuneker, 2017, p. 199)

In these examples, the linguistic context of 'he must be home' singles out either deontic or epistemic use, but annotating natural-language data, one is not always so fortunate, and 'he must be home' may very well be the complete observation to be coded for type of modality. This means that, even when using a bottom-up approach as argued for in section 4.3, this problems needs to be dealt with.

Spooren and Degand suggest low reliability scores for the annotation of coherence relations may be the result of a number of problems. First, disagreement can be a result of ambiguity, as language underspecifies meaning and context guides interpretation, as we have seen in detail in our discussions of implicatures. Second, disagreement can be a result of coding error. Ambiguity as a result of linguistic underspecification puts 'perfect agreement' out of reach, whereas the second type of disagreement should tell us 'something about the stability of our coding scheme and the theoretical conclusions that can be

drawn from our analysis' (Spooren & Degand, 2010, p. 251). As this study strives aims at maximally reliable annotation of data, as they form the input for further analysis, the problem must be dealt with in a systematic way. Therefore, five steps were taken to reach for maximum reliability of annotation. Note that a second annotator was asked to a aid in annotation, which was vital for a number of steps discussed below.[23]

The first step in reaching maximally reliable annotations of the features distilled from the literature was writing clear annotation guidelines for each feature. Each feature received a general description, criteria for classification, and codes for the actual labels to be applied. Furthermore, they were accompanied by examples. The guidelines were discussed with the second annotator before annotation began. For transparency and future use, they can be found in Appendix A. Because, as was discussed above, natural-language data tend to be more 'messy' than textbook examples, and no complete inventory of possible feature values was available for most features, the guidelines were fine-tuned by both annotators during the process.

The second step was to include not only the conditional sentences in the corpus, but also their adjacent sentences. Given the number of sentences and features in the main corpus, this context was limited to one sentence preceding, and one sentence following the conditional sentence. In most cases, this provided sufficient context to annotate context-dependent features, but I admit that it is, given the complex nature of natural language data, limited. I have tried, however, to balance the need for detailed analysis and the need for a large number of conditionals.

The third step was to include comments to observations when in doubt. Sometimes a feature may receive multiple interpretations, as was discussed above. In such cases, one value was chosen, and the considerations were included in the comments column in the corpus.

The fourth step was to randomly select a subset of 10 percent of all *als*-conditionals in the corpus. This sample of approximately 500 sentences was annotated for all features independently by both annotators. The annotations were then subjected to measurements of inter-rater agreement. In the table below the percentage of agreement is reported, as is Cohen's Kappa (Cohen, 1960). The latter is included, because it is the most used measurement of agreement, making these results ready for comparison to other annotation studies. However, Cohen's Kappa does not correct for the influence of features with disproportionately frequent values, i.e., the problem of so-called *category prevalence*. We will deal with this in the next section shortly. The results of this systematic assessment of annotation reliability are presented and discussed in the next section.

[23](Then) MA student in Linguistics M. P. M. Bogaards was found willing to carry out annotation tasks as as part of a research internship in the project.

The last step to maximise the reliability of annotation follows up on the suggestion in section 4.2, and was to select the cases of disagreement between annotators and discuss them in detail. Please note here that this was done *after* calculating the reported inter-rater agreement scores. In most cases, these discussions led to agreement. However, as specific cases of disagreement often shed light on the ambiguities that are part of natural language, they are discussed in some detail in the sections reporting on individual features in chapter 5. The motivation for this discussion is to see which proportion of disagreement was due to mistakes like mislabelling, and which disagreements suggested an actual, systematic difference of opinion of an ambiguous case (cf. the difference mentioned by Spooren & Degand, 2010)). After discussion, these insights into systematic differences and agreed upon annotations were used to improve the annotations in the main corpus.

4.5.3 Calculation of agreement

In this section, I discuss the indices of reliability used by means of the calculation of agreement between annotators. I will present and briefly discuss the results of these calculations, whereas a detailed discussion of disagreements per feature is postponed until next chapter.

As mentioned above, the simplest way of calculating and reporting the level of agreement between annotators is to use the percentage of cases in which they agree. The use of raw percentages as indices of agreement is heavily debated, however (see e.g., Banerjee et al., 1999). On the one hand, raw percentages provide easily interpretable measures of agreement between annotators, and therefore they are included in the table below, but on the other hand, it does not take into account that agreement can be reached by chance (cf. Cohen, 1960), in which case chance correlates with the number of categories available (i.e., the lower the number of categories, the higher the chance on agreement). Cohen's Kappa (Cohen, 1960), and variations such as Fleiss' Kappa (Fleiss & Cohen, 1973), correct for chance agreement, but do not take into account asymmetries in frequency distributions within features. When one category is more prevalent than others, this could lead to high agreement but a low Kappa (Gwet, 2008, p. 33). Therefore, agreement coefficients in the form of Gwet's *AC1* (Gwet, 2014) were calculated.[24] Because of the interpretability of percent agreement and the widespread use of Kappa in many research fields, these measurements are also reported below.

Gwet's *AC1* was used for assessment, as it explicitly corrects for trait prevalence (Gwet, 2008; Gwet, 2014, pp. 59–60; see also the paradoxes discussed in Feinstein & Cicchetti, 1990; Cicchetti & Feinstein, 1990). While other features have prevalent categories too (for example, an overwhelming majority of clauses in conditionals has simple present verb tense, see section 5.4), we will look briefly at sentence type. The percent agreement for this feature is 0.93,

[24]Krippendorf's Alpha (Krippendorff, 2004; Hayes & Krippendorff, 2007) were not included. For a detailed discussion, see Gwet (2011).

whereas Cohen's Kappa is only 0.72. The reason for this is that declarative sentences, as one might expect, are much more frequent than any of the other sentence types. This consequently impacts the probability of chance agreement. Gwet's *AC1* coefficient corrects for this and results in 0.92. The most extreme difference can be observed when looking at focus particles, with 93 percent agreement, but a Cohen's Kappa value of only 0.57, which is partly due to choices in coding of this variable. Therefore, a brief discussion of so-called non-necessary features is in order.

4.5.4 Non-necessary features and missing values

In principle, every conditional, apart from the insubordinate cases, has a consequent and thus a sentence type of that consequent. The classification of consequents into sentence types is both mutually exclusive, as each consequent is of one sentence type only, and exhaustive, as the four sentence types discussed in section 5.8 cover all possibilities. This is not the case for, for instance, focus particles, because not all conditionals are accompanied by a focus particle. In fact, only a minority of conditionals is. The question then is how to annotate the cases without a focus particle.

Two options are available. First, we could treat these cases as missing values and code them accordingly as 'NA'.[25] If both annotators agree on this for a particular sentence, the sentence is basically ignored in the calculation of agreement (i.e., 'pairwise deletion', see e.g., Peugh & Enders, 2004; de Raadt et al., 2019). However, conceptually, one could argue that these annotations are not missing data, or data that could not be collected, but data indicating that there was an absence of the feature, which could be argued to be a category in itself. To be clear, 'missing data' are defined in the literature on reliability measures and imputation of data as the results of situations in which 'some observers do not attend to all recording units' (Krippendorff, 2004, p. 222) and 'data are considered missing if one or both ratings of a unit are missing' (de Raadt et al., 2019, p. 559; see also Enders, 2010, chapter 1). As this is not the case here, conditionals without a focus particle were annotated for that feature using the value 'no' instead of 'NA', i.e., 'units with only one missing rating are considered and treated as disagreements, whereas units with two missing ratings are treated as agreements' ('regular category kappa' de Raadt et al., 2019, p. 564; see also Strijbos & Stahl, 2007). As Strijbos and Stahl (2007; cited in de Raadt et al., 2019, p. 560) show, different ways of dealing with missing data can produce very different agreement scores. As a result of using the 'regular category kappa' strategy, 'no' was a highly prevalent trait for the focus particle data.[26] Using *AC1* corrects for this, whereas Cohen's Kappa does not. Because, as mentioned above, the different ways of dealing

[25]'NA' stands for either 'not available', i.e., the feature exists in a given case, but is has not been annotated, or 'not applicable', i.e., the feature does not exist in a given case.

[26]The use of this strategy was also discussed with Matthijs J. Warrens (p.c.), the corresponding author of de Raadt et al. (2019).

with missing data may lead to different reliability assessments, I will include the results of both strategies in the table below. However, I note here that, while highly prevalent traits and binary coding into 'present' and 'absent' categories are discussed at length in the statistical literature on reliability assessment, the specific situation at hand is, to my knowledge, not discussed in the literature on either reliability assessment or other corpus linguistic studies.[27] Therefore, I decided to include the agreement scores for both strategies dealing with missing data (i.e., regular category kappa, pairwise deletion) for non-necessary features, which, in this study, are modality, negation, and focus particles.

4.5.5 Results of agreement calculations

The results from the agreement calculations are presented in the table below, followed by a short, general discussion. Detailed discussions are provided in each feature's section. If a feature is accompanied by '(a, c)', this means that the feature was annotated for both the antecedent and the consequent. If the feature is accompanied by '(c)' only, this means the feature is only applicable to the consequent. Lastly, if the feature is not followed by parentheses, this means that the feature is annotated for the conditional as a whole. This convention is followed throughout the remainder of this dissertation.

Table 4.6:
Inter-annotator agreement scores per feature

Feature	%	Cohen's κ	*AC1*
Clause order	88	0.79	0.86
Syntactic integration	88	0.85	0.87
Verb tense (a, c)	95, 91	0.82, 0.78	0.94, 0.90
Modality (a, c)	83, 91	0.79, 0.82	0.94, 0.89
	67, 73	*0.53, 0.62*	*0.60, 0.68*
Aspect (a, c)	79, 74	0.70, 0.65	0.75, 0.69
Person & number (a, c)	94, 86	0.92, 0.82	0.93, 0.84
Sentence type (c)	93	0.72	0.92
Negation (a, c)	93, 93	0.81, 0.85	0.92, 0.92
	73, 78	*0.59, 0.66*	*0.65, 0.72*
Focus particles	95	0.65	0.95
	49	*0.45*	*0.46*

Note. Italics indicate pairwise deletion scores.

[27] This was also discussed with Stefan Th. Gries during the Summer Institute of the Linguistic Society of America (LSA; personal communication, July 2, 2019).

Interpreting the figures in Table 4.6 along the lines of Landis and Koch (1977), all features reached substantial (0.61-0.80) to almost perfect (0.81-1.00) agreement.[28] This is somewhat surprising in two ways.[29]

First, I expected certain features, such as clause order, to reach almost 100 percent agreement (not necessarily corresponding to an equally high *AC1*, given distributions of feature values). After all, such a feature was not considered interpretative, but objectively classifiable. Although I will postpone more detailed discussion of this feature until the next chapter (see section 5.2), the main reason for the lower outcome is that a sentence such as in one in (34) below can be either classified as sentence-initial, focusing on *als* as the starting point of the conditional, or sentence-medial, focusing on the conditional as intercalated in the subordinated clause (see also the discussion in Reuneker, 2016).

(34) Ja maar ik neem wel aan dat jij als je naar Spanje gaat dat je dan al Spaans kent. (fn007887)
Yes, but I assume that you if you go to Spain that you already know Spanish then.

The decision made in this case was to regard this example as a sentence-medial case, because the conditional clause is inserted between the subject *jij* 'you' and predicate *je als Spaans kent* 'you already know Spanish', and because we see resumptive *dat* 'that' after the conditional clause and, finally, because the main clause has a verb-final word order typical for subordinated clauses, but not for main clauses of conditionals. (For a more detailed discussion of such cases, see section 5.2.)

Second, even a highly interpretative feature like *modality* scores *AC1* values of 0.94 and 0.89. This cannot be due only to prevalence of the 'no' category, as the 'uncorrected' Kappa is high too. Also notice the relatively high scores on the pairwise deletion strategy for a number of features in Table 4.6. During the post-annotation discussion between annotators it indeed seemed to be the case that in most cases, the annotators agreed on the type of modality of the clause and the motivation behind that classification. The lowest agreement was reached for (lexical) aspect, both in the antecedent and the consequent. This probably reflects the complex and interpretative nature of this feature (see section 5.6).

[28] There is criticism on using these boundaries. However, as Landis and Koch (1977, p. 165) remark, 'although these divisions are clearly arbitrary, they do provide useful "benchmarks"' for the example they are discussing. I'm using these figures in the same vein here.

[29] Also note the substantial difference between the regular category scores and pairwise deletion scores for focus particles. This is due to the low number of focus particles in general, which, as discussed above, increases the impact of disagreements.

4.5.6 Conclusion

In this section, I argued for the necessity of maximising the reliability of corpus annotation, and I suggested multiple steps before, during and after the annotation process, with a focus on chance- and distribution-corrected measurement of inter-annotator agreement. Before discussing the actual features and their distributions, I will offer an account of how the distributions of features and their associations to the dimensions *mode* and *register* are analysed in the following section.

4.6 Data analysis

4.6.1 Introduction

Before discussing the distributions of each individual feature in the next chapter, I will discuss the analysis and presentation of the data, in order to prevent redundancy by doing so for each individual feature. Although the data for each feature differ, the analysis thereof follows the same steps and assessments. These will be discussed below.

4.6.2 Data presentation

All (multi-level) features are compared on two dimensions, namely mode (spoken, written) and register (formal, informal). As there are multi-level features and two dimensions, the tables presenting these distributions of features tend to become large and complex. Therefore, I used the 'division of the visual processing of graphical displays into pattern perception and table look-up' by Cleveland (1993) to present the distributions visually for overview, while offering a more detailed view of the data by means of tables in Appendix B. For each feature, a reference to the respective section in the aforementioned Appendix will be provided.

The features will be analysed individually first in chapter 5 and explored collectively in chapter 6. The reason for doing so is that the first step allows for a detailed account of each feature, including a discussion of the literature on that feature, and an inspection of its distribution over the dimensions of mode and register. However, these features are part of the conditional constructions under discussion, and they do not occur in isolation. This means that a univariate analysis alone will not do. After all, we want to know how these features work together in interaction to give rise to implicatures of unassertiveness and connectedness. In the next section, I will discuss the univariate analysis of the data, while the specification of the (multivariate) cluster analysis as introduced in section 4.3 is postponed until chapter 6.

4.6.3 Analysis of individual feature distributions

As all features are categorical variables, and the data for which they are annotated are the same across features, the setup of these tests is the same throughout the next chapter. For each feature, its distribution over mode and register is presented. As each feature may involve associations to mode and register, more than two variables are involved in testing these associations. A simple goodness-of-fit test, such as the well-known chi-square test, will not suffice, as this would only account for main effects between two variables only. We are interested not only in associations between two variables, but in associations between more than two variables, including their higher-level associations or *interactions*. Therefore loglinear analysis was used to analyse the data (see Agresti, 2007, pp. 204–243), which is a multidimensional extension of the chi-square test. This non-parametric type of analysis is 'regarded as the method of choice for analysing multidimensional contingency tables' summarising categorical data (McEvoy & Richards, 2001, p. 867). Loglinear analysis constitutes a modelling approach, which means that its objective is to find a parsimonious model that fits the data best. As such, loglinear models combine evaluation of the fit between observed and expected cell counts with testing of main and interaction effects. This approach is also referred to as 'ANOVA for categorical data' (Scheepers, 2017, p. 887).

In the next chapter, we will use loglinear analysis to try and explain the data by finding the smallest set of variables and their interactions that estimate the distributions of the feature of interest (for an introduction to loglinear analysis, see Everitt, 1977; Kuroda, 2007).[30] In order to arrive at the most parsimonious model, backward elimination was carried out (see e.g., Howitt & Cramer, 2008, chapters 38, 39; Kuroda, 2007, p. 115; Desarbo & Hildebrand, 1980, pp. 45–46), which means that for each of the features, the full (saturated) model formed the starting point of analysis. This model always perfectly fits the data, but in most cases, it is unnecessarily complex. Therefore, components of the model were removed subsequently, starting from the highest-level interactions, until the model reached a significantly worse fit to the data. The last model with a non-significant difference to the actual data is the model chosen for further inspection by breaking down the higher-order effects (McEvoy & Richards, 2001, p. 869). Note that, like the majority of models constructed using loglinear analysis, the models in this study are hierarchical, meaning that a model including a higher-order interaction also contains the lower-order interactions and main effect of that interaction (see e.g., Desarbo & Hildebrand, 1980, p. 43). In case of significant higher-order associations (in this case, two-way and three-way interactions), the effects were broken down using separate chi-square tests (Field, Miles & Field, 2012, p. 850). In case of significant associations, a measure of strength of association was calculated, because the significance of an associ-

[30] Although loglinear analysis is seen as the categorical variant of analysis of variance for continuous data (ANOVA), please note that no distinction between dependent and independent variables is made in loglinear analysis.

ation does not tell the strength of the association (cf. Acock & Stavig, 1979, p. 1381). In other words, a significant association between for instance clause order and mode (spoken vs.written text) does not tell us what the size of this effect is. Therefore, Cramér's V (Cohen, 1992) was calculated as a measure of strength of association.[31,32]

As may be expected from large samples and multiple variables, many associations and interactions turn out to be statistically significant. As models resulting from loglinear analysis may involve complex interactions, they are not always easily interpreted. Therefore, in breaking down the higher-order effects, I found it insightful to evaluate which frequencies contributed significantly to the overall association found. One way to do this, is to perform post-hoc tests on all comparisons in the main distribution, which comes down to generating and testing each of the (broken-down) 2x2 tables. The resulting p values then need to be evaluated using the Bonferroni correction (see Harris, 2001, pp. 13–41 and Cabin and Mitchell, 2000 for a critical discussion of this correction). This correction comes down to dividing the standard alpha level α of 0.05 by the number of comparisons, resulting in a new, lower alpha level α', as shown below in (35).

(35) $\alpha' = 1 - (1 - \alpha)^{1/k}$

Only those distributions that resulted in p values below $\alpha'$are considered to be associated significantly to the dimension in question. Despite the apparent usefulness of such post-hoc testing, the results of these tests, especially for large tables, are not always readily interpretable in relation to the main features discussed. The reason for this is that all levels of dimensions are tested against each other individually, and not against the rest of the distribution. Furthermore, the Bonferroni correction is considered too conservative by some scholars (see e.g., Gries, 2013, pp. 273–274). Therefore, I chose to use the standardised residuals from the chi-square test instead (see Agresti, 2007, p. 87), which provide information on the extent each cell contributes to the significant outcome of the omnibus test. These residuals reflect the ratio of the difference between the observed and expected frequency to the standard deviation of the expected frequency, and are comparable to z-scores (see Field, Miles & Field, 2012, p. 826),

[31] In many cases in this study the tables are larger than a 2x2 contingency table because, for instance, a feature like verb tense may take four verb tenses as values. Cramér's V takes the $\mathcal{X}^2$value and divides it by the number of observations N multiplied by k-1, where k is the lowest number of categories (either rows or columns in the contingency table). As k is variable, this formula can be used for contingency tables of sizes exceeding 2x2.

[32] The following value ranges (Cohen, 1988, pp. 79–80) are used here to evaluate effect size.
1 degree of freedom: >=0.10, small; >=0.30, medium; >=0.50, large.
2 degrees of freedom: >=0.07, small; >=0.21, medium; >=0.35, large.
3 degrees of freedom: >=0.06, small; >=0.17, medium; >=0.29, large.
Although Cohen (1988) does not provide guidelines for df>3, these can be calculated by dividing the df=1 thresholds by the square root of the desired degrees of freedom, resulting in the following guidelines (see also Kim, 2017, p. 154).
4 degrees of freedom: >=0.05, small; >=0.15, medium; >=0.25, large.
5 degrees of freedom: >=0.04, small; >=0.13, medium; >=0.22, large.

meaning that they are a measure of how significant the contribution of each cell of a table is with respect to the overall chi-square value. A standardised residual of 0 would mean that the frequency of the corresponding cell does not deviate from what was expected based on the overall distribution, in turn contributing nothing to the chi-square value. The stronger the standardised residual deviates from 0, the greater the contribution of that cell to the chi-square value (see e.g., Delucchi, 1976, p. 314; Agresti, 2007, p. 38; Sharpe, 2015, p. 2). A standardised residual outside ± 1.96 is significant at $p<0.05$, a value outside ± 2.58 is significant at $p<0.01$, and a value outside ± 3.29 is significant at $p<0.001$ (cf. Field, Miles and Field, 2012, pp. 825–826; see also Sharpe, 2015, p. 3 for discussion on Bonferroni correction of these alpha levels). In other words, these values tell us whether the cell of a table contributes to the chi-square value, and if so, whether it is a weak or major contributor.

4.6.4 Conclusion

In this section, I explained how comparisons between distributions on the dimensions *mode* and *register* will be presented and analysed. As the distribution of each feature will be compared on two dimensions (mode, register), loglinear analysis will be used in the next chapter, because there may be interactions between these dimensions and features. I have also discussed the general approach to breaking down high-order effects by testing multiple lower-level associations in the final models and using standardised residuals to interpret the direction and strength of the associations found.

4.7 Conclusion

In this chapter, we first discussed the reliability of annotating types of conditionals in corpus data. The results showed that reliability was low, and the ramifications of this finding led to the choice for a bottom-up approach to conditionals, and more specifically, the clustering of grammatical features to inspect their relations to implicatures of unassertiveness and connectedness. I introduced the analyses that will be used to investigate the individual distributions of features, while a detailed account of the cluster analyses on the collective feature set was postponed until chapter 6.

As annotated features form the input of further analyses in this study, the construction of a representative and balanced corpus was discussed, and with it, the choice for a language-specific corpus study of Dutch conditionals. I also discussed the need for, and construction of a representative and balanced corpus. Before the final sampling frame was presented, the identification of the conditional use of the conjunction *als* 'if' was discussed, as it strongly determined which sentences were included in the corpus of Dutch conditionals. Next, I discussed several measures taken to ensure a high level of reliability of the manual annotation of corpus data. This resulted in annotation guidelines, double-blind

and independent annotation of a subset of the data, measurements of inter-rater agreement and post-annotation discussion. I also reviewed the results of inter-agreement calculations on the annotations in general, and postponed their detailed discussion per feature until next chapter.

Finally, I described the data presentation and (quantitative) analysis. This enables us to use the general setup for each individual feature in the next chapter, in order to get a detailed view of how the features are distributed over the parts of the aforementioned corpus. With these preliminaries set, we are ready to discuss each of the features related to specific implicatures of unassertiveness and connectedness inventoried in chapter 3 in the following chapter.