Friedrich-Schiller-Universität Jena

Philosophische Fakultät

Institut für Anglistik und Amerikanistik

Translation Effects on Impersonalization Strategies

A Corpus-based Comparison of English and German

Bachelorarbeit zur Erlangung des akademischen Grades

Bachelor of Arts (B.A.)

vorgelegt von Alexander Rauhut

geboren am 16.10.1990 in Jena

Erstgutachter: Dr. Florian Haas

Zweitgutachter: Prof. Dr. Volker Gast

Jena, den 27.08.2013

Abstract

In der vorliegenden cross-linguistischen Arbeit werden die Übersetzungsäquivalente des deutschen Pronomens *man* auf deren inhärente semantische Eigenschaften untersucht und mithilfe der empirischen Methode der Korpusanalyse quantitativ mit dem Deutschen verglichen. Das Konzept Unpersönlichkeit (Impersonalization) wird zuerst theoretisch innerhalb eines kognitiv-linguistischen Konzepts verortet und aus einer agens-defokussierenden Perspektive beschrieben. Anschließend werden zwei Lesarten des Pronomens *man* identifiziert: Die universelle und die existentielle Lesart. Des Weiteren wird die Pronomen-Eigenschaft Inklusion/Exklusion eingeführt. Anhand derer werden dann die englischen Äquivalente, die zuvor aus zwei Korpusstichproben extrahiert worden, analysiert. Die Äquivalente umfassen Personalpronomen in unpersönlicher und deiktischer/anaphorischer Verwendung, andere Arten von Pronomen (insbesondere das Indefinitpronomen *one*), Passiva, Infinitive, sowie Nominalisierungen und andere Nominalphrasen. In einer quantitativen Korpusstudie werden anschließend innerhalb zweier Korpora – dem Korpus des europäischen Parlaments ‚Europarl' und dem ‚OpenSubtitles'-Korpus – zuerst die Äquivalente quantifiziert und folgend die deutschen Sätze nach universellem oder existentiellem *man* kodiert. Hierbei werden die Satzkontexte weitestgehend ignoriert, da der Fokus auf den pronomeneigenen Eigenschaften liegt. Die englischen Übersetzungsäquivalente dienen gleichzeitig als Indikatoren. Es wird angenommen, dass es signifikante Unterschiede in der Bedeutung der Konstruktionen beider Sprachen gibt. Diese Annahme kann in der statistischen Auswertung bestätigt werden und die Pronomeneigenschaften werden im Detail untersucht. Es ergeben sich außerdem signifikante Unterschiede zwischen beiden Korpora. Die Bedeutung der Pronomen im OpenSubtitle-Korpus scheint stabiler zu sein als im Europarl-Korpus. Zum Schluss werden Probleme der Methode im Hinblick auf konzeptuelles Design (besonders die Unterscheidung zwischen interner und externer Perspektive vs. exklusive und inklusive Referenz) und die quantitative Methode (besonders die Stichprobengröße und Zusammensetzung der Stichprobe) diskutiert.

Table of Contents

List of Figures and Tables

List of Figures and Tables

1. Introduction

Man as an impersonal pronoun covers a large variety of meanings in specific contexts. It is more flexibly used than any similar construction in English. While English had had a *man*-pronoun up to Middle English times (cf. Lehmann 1995), it then lost an impersonal pronoun that covers all areas of German *man*. Hence, other strategies are used in Modern English to compensate for the lack of *man*. For example, the Langenscheidt bilingual dictionary lists four different meanings of *man* and as many as ten translation strategies (including *you, one, we, gerundial and passive-constructions, somebody, someone, they, people* and *imperatives*) (Langenscheidt 2004). In parallel corpora, all of these can actually be observed as equivalents of *man*. Additionally, there are a lot of other nouns and forms of nominalizations apart from the gerund, while indefinite pronouns are extremely rare. Since both systems do not match, there is the possibility of a regular meaning shift in bilingual situations. By using *man,* the speakers usually want to avoid direct mention of themselves or others. The interpretation of *man* is context-dependent; i.e. the referent or group of referents can only be recovered from context clues in discourse. Hence, it is considered to be non-referential (cf. Siewierska & Papastathi 2011, Gast & van der Auwera 2013). It potentially allows for interpretations ranging from *I* and *we* to *they* or any other human individual or group. English equivalents usually carry either more or less information. In the case of indefinite pronouns and personal pronouns, the potential referent is more restricted or even definite, and in nominalizations the status of the referent or the agent is completely different. Passives normally do not restrict the reference to humans, and sometimes there is even no typical agent at all, and the referent is an event or other abstract entity.

From a communicative perspective, *man* is an avoidance strategy, especially in delicate settings as in the proceedings of the European Parliament. Direct address is often risky in politics. This is, of course, also true for English speakers, who do not have the convenience of a broad, ambiguous *man*-term. Other impersonal strategies are used instead. The question is, therefore, whether German and English impersonals match sufficiently, or whether there are systematic differences in meaning. In German, *man* is often used instead of *I* or *we*, while in English *we* is a major impersonal strategy itself. Still,

there is a subtle difference between them. When, for example, someone's responsibility is in question, this subtle difference could be crucial. The following example was produced in the European Parliament. Both the English and the German sentences are translations from French *on*.

> (1) Von jenem Zeitpunkt an, nannte man die Deutschen nicht mehr 'Boches', man gab ihnen andere Namen, man betrachtete sie mit anderen Augen.
> (2) From that time on, we have not called the Germans 'Boches'; we have given them different names and looked at them through different eyes. (Europarl, 16201[1])

Both translations are identical except for the subject pronoun. The German interpreter does not use *wir*, which can be used impersonally, too. In (1), by the choice of *we*, the speaker identifies himself as being part of the group of people who "have not called the Germans 'Boches'". However, German *man* abstracts even further. Both exclusive and inclusive readings are possible, i.e. the speaker does not specify whether or not he is part of this group.

An English speaker or interpreter is forced to make a choice. As the typology developed by van der Auwera et al. (2012) and Gast & van der Auwera (2013) shows, the distribution of human impersonal pronouns in English is restricted by context and the features of the pronoun. Every strategy carries its own semantic substance, which may differ from that of German *man*. Therefore, a side-effect of the contextual restrictions is that where German would have *man* in one place English would necessarily have another perhaps more specific pronoun. Based on the typology of van der Auwera et al. (2012), I will investigate the semantic differences between English impersonal strategies and German *man*. In order to really capture the semantic shift, I will disregard the context in this paper and focus on the inherent features of each strategy (cf. features of HIPs in van der Auwera et al. 2012). The translation process itself will not play a role.

[1] The search interface of OPUS was used to extract the data (see Tiedemann 2009 for more information). The IDs after the examples indicate the position in the output of the search query and at the same time the ID in the data collection in the appendix (on CD). For the OpenSub examples, the film is cited, too.

Firstly, I will discuss the phenomenon of impersonalization in section 2 and embed the study in a cognitive linguistic approach taking the position of an agent-defocussing view. Then I will characterize the German pronoun *man* in section 3, introducing the concepts of quantification and inclusion/exclusion. In section 4, I will shortly deal with translation as a process and the use of the term *translation* in this study. In section 5, I will attempt at a *man*-centred semantic analysis of the English strategies found next to *man* in the European Parliament proceedings parallel corpus (Europarl) and the OpenSubtitles corpus (OpenSub), and I will present first data. In section 6, I will derive hypotheses. In the methodology part in section 7, the corpora will be introduced and the method of corpus analysis explained. After that the hypotheses are tested and discussed with the help univariate and bivariate statistics in section 8. Section 9 concludes and discusses the strengths and weaknesses of the empirical design.

2. Impersonalization

Impersonalization as a concept in linguistics has received attention in many different ways. A definition of this phenomenon is, therefore, highly theory-dependent. Impersonal constructions range from pronominal generics or indefinites to passives and even nominal subjects that do not denote humans (e.g. nominalization of events) (cf. Siewierska 2008). In diachronic linguistics impersonalization is to be understood as the process of grammaticalization through which a (personal) construction shifts towards an impersonal construction (e.g. German noun *Mann* → impersonal pronoun *man*). In this corpus study, however, the focus will be on the synchronic perspective.

2.1 The Formal vs. Functional and Cognitive Perspective

Siewierska discusses two different views on impersonalization: the subject-centred and agent-centred view (also referred to as instigator-centred) (cf. Siewierska 2008). Both views can roughly be associated with structural and cognitive approaches to grammar. From a subject-centred perspective, constructions that have a non-referential, expletive or non-overt subject can be seen as impersonal constructions (cf. Siewierska 2008: 116). Especially studies in the framework of generative grammar have dealt with impersonalization with

respect to a subject-centred view. These elaborations base on the idea of syntactic zeros in Chomsky's Binding Theory and the extensions in his Minimalist Theory (see Chomsky 1995). Bhatt & Izvorski (1997) and Bhatt & Pancheva (2005) deal with the external or implicit arguments of gerunds and infinitivals, proposing that the subject position is filled with a non-overt PRO. Mendikoetxea (2008) even analyses pronominal impersonals like French *on* and German *man* as being expletive (merely a syntactic place filler with little or no semantic substance), thus leaving a subject position empty. These analyses depend on the notion of the (non-)canonical subject. Pronominal impersonals like German *man* and English *you* fall under the category of non-referential subjects. The status of other English strategies such as the personal passive, infinitives and nominalizations is more complicated from a subject-centred view if not that they are not to be treated as impersonals, at all. Blevins (2003), for example, excludes personal passives from his classification entirely (see section 5.3.1 for discussion).

The agent-centred view, on the other hand, comprises constructions which feature a non-elaborated or under-elaborated agent; the agent is defocused. In the sense of cognitive grammar, the agent here is a cognitive archetype in events that prototypically establish agent and patient roles (Langacker 1991: 224). Defocussing, therefore, means that the salience of this archetype is diminished, i.e. the agent is not further specified and/or the subject cannot clearly be attached to an agent (demotion) (Siewierska 2008: 121). English translation equivalents of *man* typically carry one of these features. Summing up, Siewierska offers a scale of impersonality based on argument structure (for more detail and examples, see Siewierska 2008: 117-120, 125):

(3) a. focal argument
b. under-elaborated argument
c. demoted obligatory argument
d. demoted optional argument
e. demoted non-argument
f. no argument
(Siewierska 2008: 126)

One crucial point of this view is that it includes personal passive constructions, which fall into category (3f). The subject-centred view does not include personal passives or infinitival clauses per se, because there is an argument with subject properties that is selected in favour of the patient or another constituent other than the agent. As will be evident, the personal passive in English is one of the main strategies for conveying impersonality, at least from an English-German cross-linguistic perspective. The agent-centred view offers a solid basis for an analysis of all these strategies that occur in corpus data relative to German *man*. It is also noteworthy that, from an agent-defocusing view on impersonalization, existentials can be seen as impersonal constructions. However, Siewierska notes that they are only peripheral impersonals (Siewierska 2008: 121). The data mirrors that analysis in that existentials occur as counter-parts of German impersonal man, but are fairly rare. The different approaches are, of course, not clear cut and have significant overlap. This is the case because they operate on different levels; subject-centred on the level of morphology and syntax, agent-centred on the level of semantics and pragmatics.

One of the major impersonalization strategies of German is the pronoun *man*. For quantitative corpus research this pronoun is ideal as a starting point in investigating the impersonal strategies of other languages. As the purpose of this paper is to capture the meaning differences between English and German impersonals, the next sections will deal with the semantics of *man* and the English strategies in question. There will be special focus on pronominal strategies.

3. The Human Impersonal Pronoun *'man'*

Turning now to the German impersonal pronoun *man* deriving from the noun meaning 'man' (cf. Lehmann 1995), one can observe a significant overlap in both agent-centred and subject-centred approaches. It lacks canonical subject properties in that it does not have a full set of agreement features (Kratzer 2000), and it defocusses the agent and is non-referential. In many respects, *man* can be regarded as a prototypical impersonal construction, at least among the pronominal strategies. In Siewierska's terminology it is part of a sub-group of impersonal constructions called R-impersonals; Gast & van der Auwera (2013) use the term Human Impersonal Pronoun (HIP). R stands for reduction in

referentiality, i.e. they feature a human non-referential subject. Siewierska does not restrict this to pronouns, however. (Siewierska 2011: 57-58). In this study, the pronoun *man* is of central importance because it serves as a starting point in the empirical method applied.

Man is polysemous and different readings can be classified on the basis of different dimensions. Kratzer (2000) observes inclusive and exclusive *man;* Zifonun (2001) identifies generic and existential (particular) readings of *man*. Similarly, Giacalone Ramat & Sansò (2007) suggest a fourfold grammaticalization cline applying the features of number, referentiality and inclusion, with which the diachrony and the synchronic distribution of *man* can be captured.

(4) species-generic → human non-referential indefinite → human referential indefinite → human referential definite
Giacalone Ramat & Sansò (2007: 98)

Contemporary German *man* arguably covers the first three nodes. It cannot be used as a human referential definite. This use is common for French *on*, which can unambiguously refer to the first person plural and has, in fact, become an alternative. Direct reference to the first person is, if at all, uncommon or even completely impossible for German *man*. (cf. Zifonun 2001: 242). Gast & van der Auwera (2013) have developed a typology including and modifying the grammaticalization cline of Giacalone Ramat & Sansò (2007) and combining it with the typological approach to 3rd person plural impersonals by Siewierska & Papastathi (2011), which was based on Cabredo Hofherr (2003, 2006). They suggest a semantic map connecting all contextual interpretations of German *man* with the domain of indefinite pronouns and some anaphoric and deictic uses of personal pronouns. As differentiating feature they add the context features episodicity and veridicality. An additional feature of the pronouns is quantification (van der Auwera et al. 2012: 9ff.). They point out that the genericness of the sentence is to be distinguished from the genericness of the pronoun (the distinction is taken from Krifka et al. 1995); therefore, 'generic' refers to the context and 'universal' to the pronoun. The other variable introduced is veridicality and refers to the truth conditions of sentences containing modal or conditional operators. Since the state of affairs is not the focus of this paper, I will not go further into detail (see van der Auwera et

al. 2012: 8f. for discussion and examples, also cf. Zwarts 1995). The question is, therefore, which semantic properties come from the pronoun *man*. Only on that basis can the semantic meaning shifts between English and German be investigated empirically.

3.1 Quantification

Focussing on the features of the pronoun, there are two distinct variations which can be identified: universal and existential (roughly 'generisch' vs. 'partikulär' in Zifonun's terminology) (cf. Zifonun 2001, van der Auwera et al. 2012). These variants are also similar to the analysis in Kitagawa & Lehrer (1990), who use vague for existential with a distinct meaning, however. The two terms originate in formal semantics and are expressions of quantification. The universal quantifier is similar to expressions like all, and any; the existential quantifier is similar to someone. For German *man* this can be considered one of the main distinctions, and as a result we get two readings. Universal *man* contrasts with existential *man*. Universal *man* is to be understood as a generalization over all human individuals (species-generic) or all members of a sub-group of humans. There is no definite or indefinite individual that can be identified. It does not matter how small the group in question is. It is conceivable that universal *man* refers to a group that only contains one individual; it would still be considered universal. Hence, it is not a matter of number.

(5) Man lebt nur einmal.
(6) Als Bundeskanzler hat man es schwer.

In contrast, existential *man* denotes human individuals or groups that are potentially identifiable, but not further specified. Embedded into context the interpretations range from the speaker's self-reference to almost anaphoric interpretations similar to *they*.

(7) Man hat mir das Fahrrad geklaut.
(8) Man gewann das Spiel 3:2.

The existential reading of *man* tends to appear with verbs in past tense, as in the examples above, and can often be associated with episodic contexts. Existential *man* in such

unambiguously episodic contexts is easy to identify. In other contexts the substitution with *jemand* can be an indicator. Existential *man* is also vague in number. In reciprocal contexts it can be unambiguously plural.

(9) Man spielte miteinander Karten.

However, normally there is no identifiable number feature for *man*. Existential impersonals can also be subdivided for definite and indefinite uses (cf. van der Auwera 2012). Normally, existential *man* is indefinite, i.e. it denotes an individual or individuals that are identifiable in context but not specified. There can be definite uses of *man* when it is used denoting collectives. Compare (10) and (11), the first meaning *someone* and the second *those who are responsible for raising taxes*):

(10) Man hat mir das Auto gestohlen.
(11) Man hat die Steuern erhöht.
(cf. van der Auwera 2012: 13)

This distinction, however, does not play a significant role for *man*. The distinction between universal and existential will be taken up in the empirical section, and both readings will be treated separately.

3.2 Inclusion/Exclusion

Another potential property of *man* is inclusion or exclusion of the speaker or hearer (cf. Dimowa 1981, Kratzer 2000, van der Auwera 2012). Dimowa (1981) comes to the conclusion that *man* has 6 pronominal sememes. Hence, *man* is supposed to have six additional readings corresponding to the personal pronouns of German. Zifonun's (2001) analysis, however, makes a clear cut between the sentence meaning and the semantics of the pronoun *man*. On that basis, she points out that speaker- and hearer- exclusivity and/or inclusivity is only generated in context. Interpretations of *man* corresponding to personal pronouns can be seen as contextually embedded generic/universal or existential *man*; therefore, they are triggered by implicatures, which are not obligatory (cf. Zifonun 2001:

241). This also affects the inclusion or exclusion of the speaker and/or hearer. The examples in Kratzer (2000), however, suggest that inclusivity is dependent on the morphosyntax of *man*.

(12) *Wenn man seine Brille aufsetzte, kriegten wir Angst. (Kratzer 2000: 5)

(13) *Wenn man seine Brille aufsetzte, kriegtest du Angst.

The speaker-inclusive reading in (12) and similarly the hearer-inclusive reading in (13) are impossible. That means that speaker and hearer inclusivity can be eliminated by possessive pronouns. Another way how this is achieved is through predicative noun phrases (Kratzer 2000: 4).

(14) *Als Hüter des Gesetzes hat man mir erklärt, ich könne hier nicht wohnen. (Kratzer 2000: 4)

These restrictions arise from the morphological and syntactical properties of *man*. Still, the view in Zifonun (2001) will be taken here since predicative noun phrases or possessive pronouns occur very infrequently in the corpus data. For this corpus analysis, the assumption is that *man* remains ambiguous for exclusion or inclusion. In the data, a definite decision can almost never be taken in favour of either inclusion or exclusion. The reason for this is again that *man* is often used as avoidance strategy, so the speaker does not want to specify whether or not s/he or the hearer is included. The only way to disambiguate inclusive and exclusive readings is, therefore, by judging from the context.

4. Translation

Before I turn to impersonal strategies in particular and to a first overview of the data, a few remarks ought to be made on the notion of translation. The very process of translation is actually of secondary importance. This paper focusses on meaning differences of impersonal strategies as if English and German sentences in the corpus were on one equal level. Hence, translation effects actually refer to the discrepancies that are a result of the two different pronominal systems of English and German. The lack of a *man*-pronoun in

English must be compensated by other means, and vice versa the existence of such a polysemous pronoun in German poses the danger of (systematic) meaning shifts in multilingual contexts, as for example in the European Parliament. The choice of Europarl and OpenSub as corpora has the advantage that the creative input of the translator or interpreter is minimized. As the translation process itself is disregarded here, it also does not matter what the source language of the utterances is. At least in Europarl, there are many different source languages in the corpus. A translation from English into German cannot be found at all. In OpenSub the source language is in most cases English. The design suggests a side by side of English impersonal strategies and German *man*. The 'translational' equivalents in the next sections are, therefore, taken to be parallel correspondences without reference to actually being a result of translation.

5. Strategies of English

Before the particular English strategies that co-occur with *man* in the corpora are looked at in more detail, I will sum up the features of *man* and give an overview of the equivalents in the two corpora. *Man* is considered here to be either universal or existential, but normally remains ambiguous with respect to exclusive or inclusive readings. It cannot be a deictic or an anaphora. The closest translations according to dictionaries are *you*, *one*, *we*, and for existential *man*, it is *they*. In the following plots one can see the frequencies of the English strategies in the sample of Europarl. Note that participles do not appear in the chart although they were previously mentioned as strategies. They were reassigned to other strategies according to the subject of the sentence (see section 7.3).

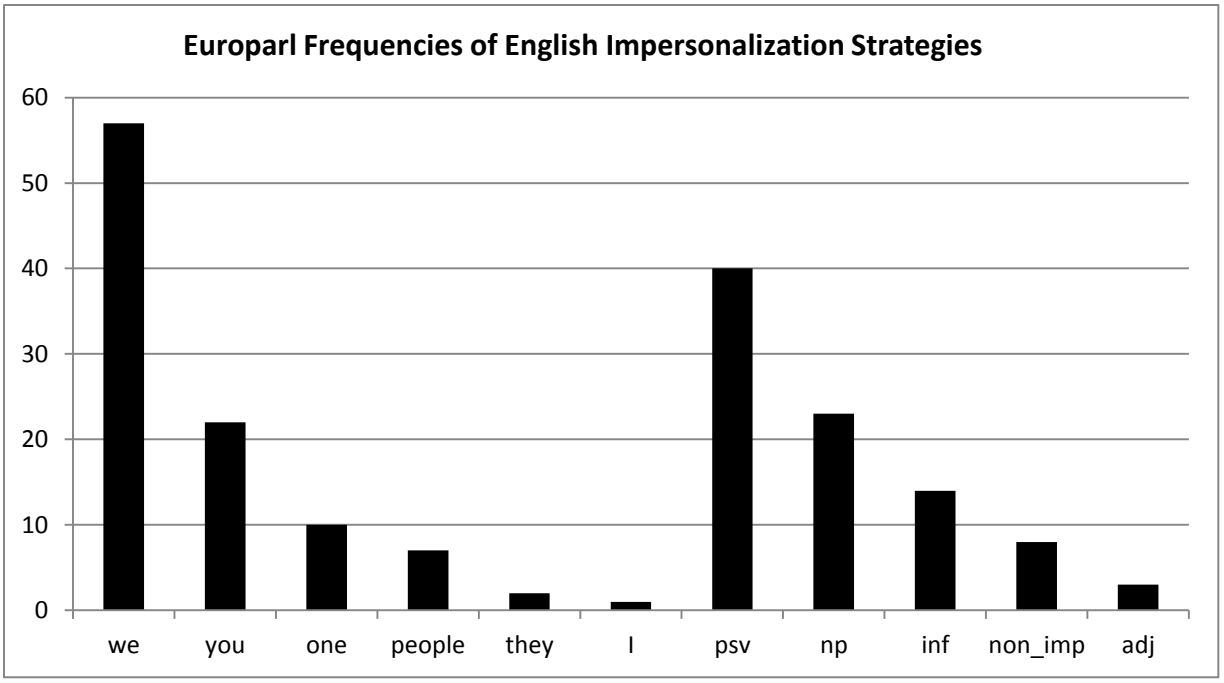


Figure 1: Europarl Frequencies of English Impersonalization Strategies (n=187)

The category 'non_imp' comprises non-impersonal uses of pronouns: anaphoric *they*, deictic *I* and *someone*. The category 'np', which stands for noun phrase, can also be sub-divided into nominalizations (especially gerunds), collective nouns (for example *parliament*) and definite descriptions. All of these can be expected to behave differently since this group is highly inhomogeneous, and especially collective nouns seem to have a special status relative to impersonals. With these sub-divisions we arrive at a total of 15 strategies (some of them were only grouped for better visibility). In the sample taken from the OpenSub corpus we get a different picture and a few more translation equivalents:

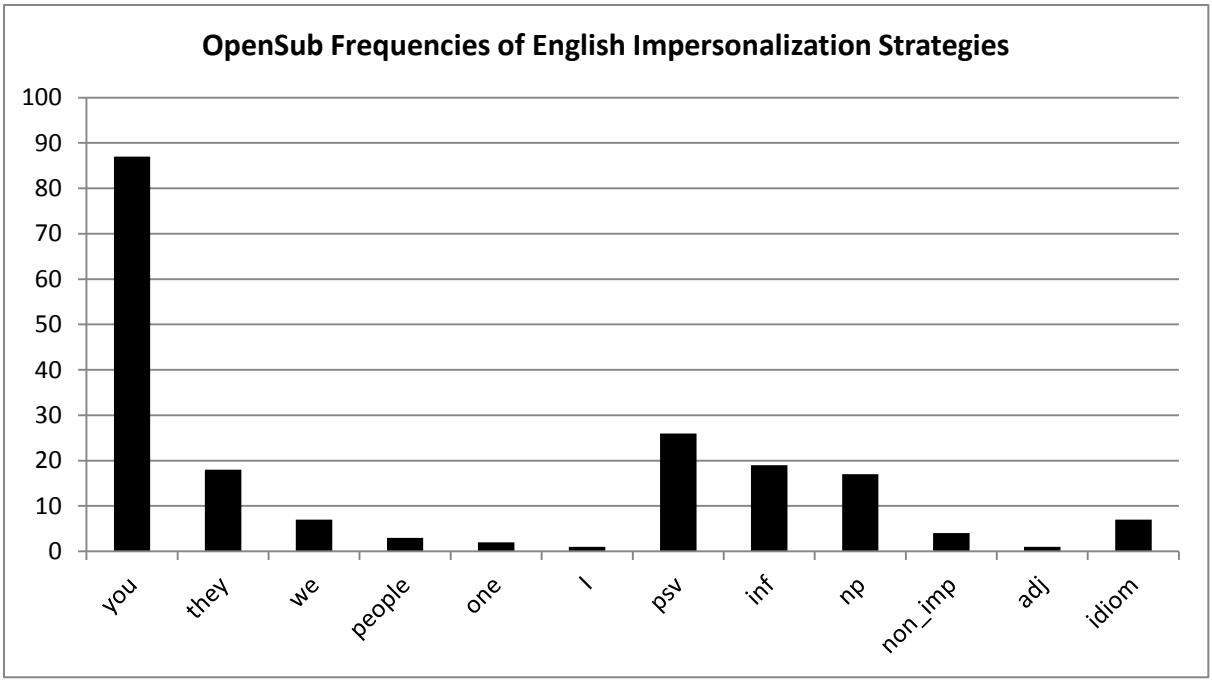


Figure 2: OpenSub Frequencies of English Impersonalization Strategies (n=192)

In contrast to the Europarl sample, we have here a category for idioms, which is the result of the more colloquial register. Those are sentences that reflect the meaning of the *man*-sentence without containing an identifiable strategy. Also, there are a couple of demonstrative pronouns, the indefinite pronoun *anybody* and in the category 'np' there are two occurrences of the generic human (*a man*, and *a guy*), which is also noteworthy. Adding these up, we get 18 strategies that have to be considered in the following sections. There will be special emphasis on pronouns.

5.1 Personal Pronouns

Uses of impersonal pronouns overlap. Some pronouns carry semantic information for certain features and others stay ambiguous, but both might co-occur in the same contexts. *Man* covers all impersonal uses of English and German pronouns as suggested by van der Auwera et al. (2012). Hence, the same semantic analysis for pronoun features will be carried out for every English strategy here. Additionally, it is interesting to look at the areas that *man* does not cover. These are deictic and anaphoric uses, and also the uses of indefinites. This section will investigate into the areas where the meaning of *man* and the meaning of the corresponding pronoun mismatch and the semantics of the pronouns will be compared.

For the decision about a pronoun's inclusion or quantification feature I stayed close to the semantic map in the typology of van der Auwera et al. (2012).

Second Person Pronouns are among the most frequent correspondences in the corpora, and they have, in fact, a predominant status in the OpenSub data. Impersonal *you* is a variant of its homonymous personal counterpart, so there is already an etymological difference to *man*. According to the semantic map proposed by van der Auwera et al. (2012), impersonal *you* cannot acquire an existential reading. Hence, it can be assumed that impersonal *you* is always universal.

(15) *You have raised the taxes. (intended: They have raised the taxes.)
(16) *You have knocked at the door. (intended: Someone has knocked at the door.)

Impersonal *you*, however, has to be distinguished from deictic uses of the personal pronoun. The personal pronoun can denote definite referents that are specified by the setting of discourse. Those are no impersonal uses; it is, however, hardly possible to code them in the data. Many if not most of the instances of *you* can receive either deictic or impersonal interpretation. Nevertheless, the meaning parameters introduced above can be applied without any problem. Concerning the feature inclusion *you* seems to be vague or at least in most cases ambiguous. Thus, it does not differ from German *man* in this respect. Although impersonal *you* is grammatically singular (Kitagawa & Lehrer: 1990: 744) because it takes the reflexive form *yourself*, there is no indication for the actual number of the potential referents since it has no existential reading, anyway. The singular agreement of impersonal *you* suggests that the 2nd person plural pronoun is restricted to deictic uses. The difference between the two is not clear cut in absence of a reflexive.

First Person Pronouns are also a major impersonal strategy in English. The plural form *we* is one of the most frequent strategies relative to *man* in Europarl. *We* always signalizes the speaker's membership of a group or at least the affinity to that group. Therefore, it is always inclusive in the sense that even if the speaker is or was not part of the action in question, s/he puts himself into the collective. In the last section of this paper, there will be a discussion of the concept of 'internal perspective', which is associated with this particular

kind of contextual exclusiveness. For now, I will stick to the analysis that *we* is always inclusive. The speaker, at least, considers himself included or is hypothetically included. The difference between deictic *we* and impersonal *we* is problematic since deictic *we* also denotes an unspecified group (cf. Kitagawa & Lehrer: 745). Whether or not that group is identifiable in discourse in the sense of a strictly deictic use remains ambiguous in most cases. Especially in Europarl, most of the instances of *we* are addressed to the whole collective of the parliament or the faction without referring to anyone in particular. This use is different from the purely impersonal *we*. Consider (17) in contrast to (18).

> (17) (…) By talking we come to understand people. (Europarl, 2201)
>
> (18) However, I would like to point out that we should also make use of our experience of managing other funds, especially the Structural Funds (…) (Europarl, 13201)

In (17) the pronoun is clearly impersonal; it is species generic. The proposition is not only true of the members of parliament. It resists the pronoun shift in reported speech and can be replaced by *one* without changing the essential meaning (Kitagawa & Lehrer 2011: 742). This is not true of (18), where the speaker addresses a vague group, identifiable but without specification (cf. Kitagawa & Lehrer 1990). The substitution test with *one* does not always produce a straightforward result. Even in (17), the deictic reading is possible as well (i.e.: 'We – I mean you and me – come to understand people by talking'). In the following, occurrences of *we* are always treated as impersonal, keeping in mind that the possibility of a deictic reading is never eliminated completely. This is not too big a problem, though, because a deictic *we* can also be classified as inclusive. With respect to quantification, there is always a universal reading. No matter how small the group is and whether or not it is definite or indefinite. *We* always generalizes over it. First person singular pronoun *I* can be used as impersonal, as well, but is mainly restricted to hypothetical contexts (Kitagawa & Lehrer 1990: 742). As a possible diagnostic, impersonal *I* can also be replaced by *one* without an essential shift in the meaning of the utterance, and it resists pronoun shift in indirect speech. Impersonal *I* is due to its nature inclusive and universal. In both samples, *I* as corresponding translation of *man* is extremely rare. Only in the Europarl sample can an occurrence of *I* be replaced by impersonal *you* or *one* without a change in meaning.

(19) This lack of transparency leads to a great many overly-dramatic, interest-led articles in the press, such as the one I read at the weekend in the Financial Times, which attacks the WTO's public image. (*Europarl*, 5801)

Third Person pronouns are another group that is proposed by dictionaries as major strategies although they are rare in both corpora. 3rd person plural *they* is typically used with existential quantification ('vague' in Kitagawa & Lehrer 2011)[2], i.e. a subset of individuals can be identified, but neither a specific individual nor the group as a whole is referred to. Very often a collective is implied, as e.g. the parliament in (20).

(20) They have raised the taxes.

When t*hey* is used as an impersonal, it lacks an antecedent in discourse unless it is an impersonal *they* itself (Siewierska & Papastathi 2011: 585). All results in the OpenSub corpus carry this feature; however, in Europarl, occurrences of *they* are extremely rare and many have an antecedent as in example (21). This makes a distinction easier than in the case of deictic pronouns.

(21) I asked the ushers to remove them; they said that they were authorized.
(Europarl, 32201)

In the example, *the ushers* is the antecedent of *they* which shows that the pronoun is in this case used anaphorically rather than impersonally. Another feature that distinguishes *they* from *man* is that it is always exclusive (cf. van der Auwera et al. 2012: 12). Although grammatically plural, *they* is vague in number (ibid. 22). Summing up, *they* differs from universal *man* in quantification, and it is always specified as exclusive. 3rd person singular pronouns in English cannot acquire impersonal interpretations as they are always

[2] Kitagawa & Lehrer (1990) restrict ‚impersonal' uses of personal pronouns to generic/universal readings. (cf. ibid.: 742) The sub-division into impersonal – vague – referential uses correlates with Siewierska's (2011) groups: (quasi)-generic – episodic – specific. 'Impersonal' and 'vague' also correlate with universal and existential.

anaphorically anchored or interpreted as a deictic. (Kitagawa & Lehrer 2011: 747) The corpus sample features no 3rd person singular personal pronouns whatsoever.

5.2 Indefinite Pronouns and other Pronouns

Another main source of impersonalization strategies is the class of indefinite pronouns. Indefinite pronouns are closely related to the domain of impersonals. *Man* is often listed as indefinite pronoun (cf. *Duden* 2013). As mentioned above, the indefinite pronouns *one*, *someone*, and *anyone* among others are common translations in dictionaries. All of them appear to be infrequent in the corpora, however. Yet, due to the size of the sample there cannot be definite judgements about that.

The pronoun *one* derives from the numeral *one* and, therefore, has to be distinguished from it, but there are no occurrences of the numeral in the corpus samples. The pronoun is often treated as high register synonym of *you*. In fact, it can always be replaced by *you* without a change in the informational content.

(22a) You could see that he was drunk.

(22b) One could see that he was drunk.

The claim that it is a near synonym of impersonal *you* is consistent with the typology in van der Auwera et al. (2012). "The uses of you are the same as those of one, except that we can also include the direct reference to the hearer, (…)" (ibid.: 23). This suggests that it has the same meaning as *you*; i.e. it is always universal and ambiguous for inclusion. Similarly to *man*, *one* is often used to refer to the 1st person. However, the same analysis is applied here that the alleged synonymy with personal pronouns is a result of conversational implicatures (see Zifonun 2001 in contrast to van der Auwera et al. 2012: 20). Speakers, so to say, 'abuse' the genericness of the pronoun to justify their own actions. Still, one can treat it as an embedded universal *one*. In contrast to *you*, there is no deictic counterpart of *one*.

The other two indefinite pronouns that appear in the sample are *anybody* and *someone*. The main difference to *man* is that they establish a referential link, but do not specify the referent. Therefore, they behave completely differently. Indefinite pronouns

might receive impersonal interpretation in conditionals (van der Auwera et al. 2012: 13)[3]. *Anybody* is similar and also functions as impersonal in conditionals. Both occurrences in the corpora can actually be interpreted as impersonals. According to the van der Auwera et al. (2012), the indefinite pronouns used as impersonals have the features inclusive and universal.

Another strategy that is commonly associated with impersonalization is the quasi-pronominal noun *people*. It is a species-generic noun. Nouns like that do not normally require an article. Hence, it can be argued that it has certain pronominal uses. When compared to *man*, it can be seen as being situated in an earlier stage of the impersonalization cline in (4). Consequently, it is more restricted with respect to impersonal uses. Most often one can find *people* used as a pronoun with speech act verbs (Gast & van der Auwera 2013: 8). The following examples are taken from the corpora for illustration.

(23) You really are quite bright, despite what people say. (OpenSub, Batman Forever, 32)

(24) (…) [P]eople could say that this House is also being hypocritical about fighting fraud. (Europarl, 27801)

With the pronominal status of *people* being unclear, anyway, it is especially difficult to distinguish it from the noun *people* as long as there are no morphosyntactic clues present (like for example articles or post-modifiers). *People* is vague for inclusion and is existential. There exists a whole group of people rather than certain individuals, but it does not generalize over all humans. It can be replaced by *they* with little difference in meaning (Gast & van der Auwera 2013: 11). I argue that this little difference lies in that inclusion is possible for *people* while necessarily excluded for *they*. The speaker does not necessarily have to be part of this group. Only if (25) is uttered by a German would *people* be inclusive, and even then it is not an obligatory reading. The implicature can be cancelled.

(25) In Germany, people eat Bratwurst (but I don't).

[3] In van der Auwera et al. (2012) they actually speak of *somebody*. To simplify matters, *somebody* and *someone* are treated as synonyms here.

Finally, the last pronouns found in the corpora are demonstratives. These are untypical equivalents and are probably a result of a rather free variation in translation since they are not used as impersonals in English, and since they do not appear in the Europarl data and are extremely rare in the ObenSub sample. They have deictic interpretation. The other parameters do not actually apply, but are treated as ambiguous here. The next section will now turn to the non-pronominal and at the same time more complicated strategies found in the corpora.

5.3 Status of Non-Pronominal Strategies

5.3.1 Passives

When comparing English and German, one major problem with the classification of impersonals is the status of subjectless passives, gerunds and infinitives (bare and to-infinitive). Without the convenience of zero-pronouns of the quality of PRO_{arb} (i.e. Bhatt & Izvorski 1997), these strategies are quite different from their pronominal counterparts as they do not necessarily imply a human referent, although they are used similarly in discourse. In (26), the agent implied is necessarily human, whereas in (27) the agent could be an animal or even a natural force.

(26) You cannot destroy this tower.
(27) This tower cannot be destroyed (by anyone / an elephant / a thunderstorm).

Very often, however, the context restricts the interpretation of passives so far that only human agents can be implied.

(28) A friendship with Saruman is not lightly thrown aside.
(OpenSub, The Lord of the Rings – The Fellowship of the Ring, 101)

The human interpretation of the agent in (28) is due to the fact that non-human entities are not likely to be friends with the exception of personifications. It is, therefore, pragmatically

implied by the concept of *friendship* in this utterance. In the corpus sample, all passives imply a human agent. They can be seen as alternatives to pronominal impersonals, and therefore, apparently qualify as impersonal strategies. At least in the case of the passive, this is a highly problematic assumption. Wales even takes the extreme opposite position, namely, treating *man/one*-constructions as a "'third person equivalent of the passive voice" (Wales 1996: 80-81). Blevins challenges this view in that he argues in his 2003 paper that the "subjectless form of a personal verb is conventionally interpreted as referring to an indefinite human agent, irrespective of the source of its subjectlessness" (Blevins 2003: 481). His point is that the logical subject is merely suppressed and not unexpressed. That again corresponds to a subject-centred view on impersonalization and contrasts with the agent-centred view as outlined above. Scrutinizing the differences between personal passives and impersonal strategies, however, would go too far for the purpose of this paper. Since the agent is suppressed, there is no entity in the sentence that could have semantic features of its own since this paper does not suggest zero items. Consequently, passives will be treated here as being vague in every respect. The (human) agent can only be recovered from the context. Arguably, in every context in the semantic map by van der Auwera et al. (2012) there can be a passive. Therefore, (the suppressed agents of) passives are treated as ambiguous for quantification and inclusion.

5.3.2 Infinitives

The next large group of pronounless constructions is that of infinitives. In English, there are two forms of the infinitive: the bare infinitive and the infinitive with the particle *to*. The first one is a hapax legomenon in the OpenSub sample and probably just the result of ellipsis. Hence, *to*-infinitives will be simply referred to as infinitives from now on. As translation equivalent in the samples, the infinitive typically appears in wh-clauses as in (31) or in combination with a dummy-*it* in subject position as in (29). Interestingly, those uses seem to pattern with certain constructions in German. In (30) and (32) the two corresponding sentences are shown.

(29) It is possible to regulate production (…)

(30) Man kann die Produktion doch auch regulieren, (…) (Europarl, 30601)

(31) You know how to do those?

(32) Weisst du, wie man die macht? (OpenSub, Robots, 188)

At first glance, this is at the same time a striking difference between Europarl and OpenSub. In the Europarl sample, with the exception of two examples, all infinitives found fit the pattern in (29) (the other two were a *there*-existential and one with the infinitival clause in subject position). On the contrary, 11 out of 18 *to*-infinitives in OpenSub followed the pattern in (31). The infinitive, similarly to the passive, usually implies a human agent (Wood 1956). In this respect, it is similar to the passive and it will, therefore, be treated likewise.

5.3.3 Nominalizations and Nouns

Finally, there is a completely inhomogeneous group of noun phrases with a head other than a pronoun. It comprises a variety of different head nouns. For the sake of brevity those constructions will be referred to simply as 'NP'. The use of such constructions probably has stylistic reasons in many cases. Still closest to the concept of impersonalization are collective nouns. Apparently, when there is a combination of a locative plus *man* in German, collective nouns are likely to appear in English.

(33) Insbesondere **in den USA** verfolgt **man** unseren Eifer mit gewissem Erstaunen und Neid.

(34) This determination is something which **the United States**, in particular, is following with a measure of astonishment and jealousy. (Europarl, 18801, emphasis added)

The meanings of both sentences are equal. The collective noun here, of course, is a metonymy. *The United States* is an abstract entity and is used instead of an unspecified group of Americans. In that sense it is universal. The noun phrase generalizes over the whole group of American citizens (or a subgroup of American politicians). The membership of the collective described by the noun is context-dependent. Another type of NPs that can be associated with impersonalization is the generic human. Phrases like *a guy* or *a man* denote

the whole species. Looking at the first node of the grammaticalization cline by Giacalone Ramat & Sansò (2007: 98), these are at the very borders of impersonalization. Naturally, those are always inclusive and universal provided the speaker and/or hearer is *a guy* or *a man*. Definite descriptions of humans can also be compared to the other impersonal strategies; they are exclusive and existential. The rest of the NPs cannot really be analysed in parallel to the pronominal strategies. There are gerunds, verbal nouns and abstract nouns. Often the verb phrase of *man* mirrors the noun phrase in the English equivalent.

(35) Dann fragt man sich (…)

(36) The next question would be (…) (Europarl, 2401)

Action nominalizations as such do not imply human agents (cf. Siewierska 2008). They abstract away further than the pronominal strategies do. Also gerunds are different from the strategies listed above. Since infinitives and gerunds are often interchangeable, one could conclude that they can be treated alike. However, Wood offers an analysis that sets the two constructions apart with regard to reference: "(...) where the infinitive, although it does not specify an agent, usually implies one, the gerund represents the activity as it were *in vacuo* without reference to any agent or occasion" (Wood 1956: 1). Hence, gerunds have to be treated differently for they express another kind of reference. With all occurring strategies characterized and all data extracted, there will be an attempt at formulating certain predictions now.

6. Hypotheses

The question posed in the very beginning of this paper was whether English impersonal strategies differ from German *man* systematically in meaning. The data from the parallel corpora can provide answers to that question. Resulting from the above discussion, there can already be made some predictions. In the following empirical analysis, the hypotheses in (37)-(40) will be tested.

(37) H_1: The meaning between English impersonal constructions and German *man* differs significantly.

(38) H_2: There are significantly more meaning shifts in Europarl than in OpenSub.

(39) H_3: *Man* tends to be disambiguated in the English equivalent concerning inclusion.

(40) H_4: English strategies with both ambiguous features for inclusion and quantification are disambiguated in German.

(37) results directly from the research question of this paper. H_2 will be tested since the register of Europarl features a lot of long and obscured sentences, and meanings often remain ambiguous and the chance is high that the English equivalent is more specific than German *man*. English does not have such a universally applicable avoidance strategy such as *man*. The last hypothesis is motivated by the observation that passives are even less specified than *man*. The prediction is that the German equivalents tend to be more specific. Some data that will be used for supporting those hypotheses has already been presented. The next section is dedicated to how the data was gathered and the major steps of operationalization will be discussed.

7. Methodology

This study is corpus-based, i.e. the data is collected with a certain aim basing on existing theory, and hypotheses are formulated beforehand. The research question being here whether there are systematic meaning differences between English and German impersonalization strategies. Nevertheless, there is a corpus-driven part, too. The collection of English strategies has been gathered prior to carrying out the actual empirical study. This is also the reason why a considerable part of the data has already been presented. This step was necessary because every thinkable English construction is a potential candidate for mirroring German *man*. This study has both a quantitative and a qualitative aspect. The qualitative portion is in the exploration of *man*-equivalents in two very different corpora, the grouping of these and the attempt of a cross-linguistic semantic analysis relative to the German pronoun. This provides the basis for the following quantitative part. The data is organised in two samples drawn from the Europarl and the OpenSub corpus, both of which contain spoken language (written as if spoken). In the following section, the two corpora will be introduced in more detail.

7.1 Choice of Corpora

Europarl is a parallel corpus collected from the proceedings of the European parliament. 11 languages are included and aligned sentence by sentence (cf. Koehn 2005). For the pairing English and German there are 1.3 million sentence pairs with over 61.5 million words[4]. The project was initiated with the aim of enhancing statistical machine translation. However, such a large parallel corpus is of particular value in linguistic typological and cross-linguistic research. The register is written as if spoken, although it is probably closer to the written one. The speeches in parliament are all prepared as written documents and revised as such. They are very consistent in style, which is high register.

OpenSub is a corpus collected from film subtitles taken from the web page *www.opensubtitles.org*. The project OPUS, which is short for *Open Source Parallel Corpus* has compiled and aligned material from 18,900 films in 59 languages (Tiedemann 2009: 2). The database is growing continuously. Currently, the alignment of German and English produces a parallel sub-corpus of 4.6 million sentences and over 53 million words. The genre is film, so the register varies. Archaic language (e.g in The Lord of the Rings) is included as well as modern slang; there is high register and low register language. Presumably, the low register is predominant. The advantage of the corpus is that it represents the spoken register. Of course, the dialogues are scripted so it is rather to be considered written as if spoken, but in exchange it is a relatively large and freely available corpus of spoken language. It is very different from the Europarl corpus, which represents higher specialized language. Already when looking at the mere frequencies of *man* in both corpora, something very interesting can be observed. Although both corpora do not differ much concerning the amount of words, Europarl features *man* 63 times more often than OpenSub (42931 matches versus 673 matches). Hence, *man* is a much more important strategy in Europarl, presumably also impersonalization.

Europarl and OpenSub as sources for this empirical study have the advantage that the translations are as close to each other as possible. This has two very different reasons, but probably the same outcome. In Europarl, the interpreters have to stick to the original as

[4] <http://opus.lingfil.uu.se/Europarl3.php>

closely as possible because in the political context even the tiniest meaning difference is dangerous during negotiations. In OpenSub the variety of translations is limited by the width of a TV-screen and partly by the lip synched voice output. The effect of this is that the length of utterances, and therefore in parts, the length of the strategy used cannot differ too much. Still occasional huge deviations can be produced by the translators; but in a quantitative analysis, these should not create systematic patterns. The restriction in the length of the output is at the same time a big weakness and strength of the corpus. The aim of this study is to find out systematic deviations of the meanings. Therefore, two different corpora can give insight into different aspects of language use. Nonetheless, it must be emphasized that the samples of the corpora cannot actually be compared and not at all mixed. Eventually, there will be two different results for both corpora.

7.2 Sample

There is, of course, still the question why this study has been *man*-centred. In the last paragraphs, other German impersonal strategies were not taken into account and the English strategies were always evaluated relative to *man*. This has practical reasons. The choice of *man* is a useful and efficient methodological procedure. *Man* is a prototypical specimen of HIPs. Being that, it has been the centre and starting point of many theoretical and typological approaches to impersonalization (i.e. Giacalone Ramat & Sansò 2007). What makes it especially interesting for studying impersonalization in other languages than German is, in fact, its written form. There are no homonyms (especially homographs) of *man* which are not HIPs. The English equivalents *you, one, people*, etc. are usually used as personal pronouns, numerals or nouns. The same is true of the other German impersonal strategies (*wir*, *die*, the reduced form *se*, *du,* impersonal passives, etc.). Carrying out a corpus study extracting those words directly from an English corpus would, therefore, be an extremely elaborate if not impossible task since every single sentence has to be scrutinized to filter HIPs. Even more difficult would it be to identify passives in untagged corpora, with the status of passives among the impersonalization strategies being problematic, anyway.

Another issue is the paradigm of *man*. Both samples only contain *man*-sentences and their counterparts. Since *man* is restricted to subject position, grammars usually suppose the

forms *einen*_acc and *einem*_dat as suppletive forms for the object positions. For both practical and theoretical reasons, I did not create a weighted sample including these forms. The oblique forms of *man* coincide with the oblique forms of the indefinite pronoun *einer* and differ in meaning systematically with respect to their syntactical position (van der Auwera et al. 2012: 25ff.). This and the homography with the numeral and the indefinite determiner are the reasons why the oblique forms of *man* are disregarded in this study.

Two samples were drawn with an approximate number of 200 sentences. In order to avoid effects from different film genres or particular films, I picked out systematically a random sample from the matches in the corpus. In total there are 673 matches in OpenSub. Every third sentence was included, which resulted in a sample of 225 sentences. In the case of the Europarl sample there were 42931 matches overall. Taking every 200th sample, I extracted 215 sentences. Randomization is necessary because the idiolect of individual speakers could otherwise be systematically overrepresented in the final statistical analysis. During the coding, some of the sentences were removed from the sample due to wrong sentence alignment or sentences that were not translated in one of the languages. Also when the meaningful part, i.e. the whole clause, containing *man* was simply left out in English, the sentence was removed. Doing this, I avoided uninterpretable null-translations. The result is a sample of 192 from OpenSub and a sample of 187 from Europarl. The remaining 379 sentences were then coded to form indicators. In the next section, the system of this coding process is documented.

7.3 Coding

Much of the actual operationalization has already been done in sections 3 and 5 by analysing the meaning components of each strategy. For the meaning of the English counter-parts the translation strategy itself serves as an indicator. Prior to that, the strategies had to be identified. There were two steps involved. The first coding of English translation equivalents was rough and concentrated on the actual grammatical form of the translation. The pronoun *man* in the German half and its verb phrase was taken as point of reference. In a second step the data was cleaned and certain constructions were reassigned to others.

Other strategies were then subdivided if necessary. In the following, the major coding decisions are listed.

Firstly, participles are not regarded as independent strategies. Participles in participial clauses can be attached to an agent in the main clause. This is a structural difference to German that predicates can be detached from the subject in adverbial clauses by participial clauses. Whenever there is a subject available in the main clause, it is coded as the translation strategy.

> (41) Da **man** diese Krankheit nicht heilen kann und ihr nicht vorbeugen wollte, (…) muß man sich heute mit der Beruhigung der Verbraucher zufriedengeben.
> (42) Not *knowing* how to cure this disease and not having had the will to prevent it (…), **we** are now reduced to bidding to reassure the consumer. (Europarl, 2601, emphasis added)

In (42) the italicised participle corresponds to German *man* in (41), but it can be attached to the subject *we* in bold. One could paraphrase (43) as follows.

> (44) We do not know how to cure this disease and do not have had the will to prevent it, and (therefore) we are now reduced to bidding to reassure the consumer.

In the case of dangling prepositions, which do not connect with the subject of the main clause, the treatment would have to be similar to that of infinitivals. Due to the nature of both source corpora the occurrence of dangling prepositions is unlikely. The reason is that English prescriptive grammar still inhibits the use of prepositions like that. Both Europarl and OpenSub are composed of revised written texts (or written as if spoken). Therefore, grammatical errors as such (from a prescriptive point of view) do not normally occur. In fact, the two samples do not feature real dangling prepositions. In a few cases, the subject in the main clause was an expletive *it* or *there*; however, in these cases, the logical subject of the sentences could be unambiguously identified as the subjects of the participle. Past participles have always been treated as passives since they are similar as it is possible to attach a by-phrase to them. There were a few *-ing*-forms that remained. All of them were

adjectives, gerunds or verbal nouns. Such forms were identified as adjectives when they were a complement of a copula. Gerunds and verbal nouns were treated likewise as nominalizations. Secondly, I distinguished anaphoric *they* (they_anaph) from impersonal uses of *they* (they_imp). As mentioned above *they* is anaphoric when preceded by a noun phrase other than another impersonal *they*. Furthermore, in the case of long passives, the subject of the *by*-phrase was taken. Imperatives were put in the category of *you* since they always address the second person in English. *There*-existentials were assigned to NPs with respect to their complements. At last, infinitives can also be attached to a pronoun as in sentences like (45) or imply one as in (46).

(45) It is nice for me to come home.
(46) What's it like to have your face on the cover of every magazine? (OpenSub, Batman Forever, 29)

In the latter example the infinitive implies *you*, which is taken up by the possessive determiner following it. These were extremely rare, which is probably due to the fact that these are equivalents of impersonal *man*.

The two different interpretations of *man,* universal and existential, were coded with the help of a substitution test. The occurrence of the pronoun was labelled universal when it could be replaced by *jeder* or *niemand* and existential when the substitution with *jemand* produced a sentence with an equivalent meaning (cf. Zifonun 2001: 240). If one of the substitutes was not possible at all, the respective reading was excluded. Only if both substitution tests did not work – for example, because there were structural restrictions – was there a further interpretation of the sentence. Hence, in many cases the result was true for universal and for existential. This reflects the ambiguity of many *man*-sentences. The senses were not disambiguated on purpose. An ambiguous sentence on the communicative level has a function of its own in the same way as a sentence that can be interpreted unambiguously. Eventually, for the parameter quantification there are three values: existential reading possible, universal reading possible and both readings possible. Clearly, universal readings of *man* strongly correlate with the presence of modal verbs or conditionals. Many of these are idiomatic and are almost used as adverbials.

(47) Man könnte sagen, dass …

(48) Wenn man bedenkt, dass …

Constructions like that do at least not really intend an existential reading. The meaning of these can be interpreted like 'everyone can say x, and what is true for everyone is true for every single human as well'. Exclusively existential readings, on the other hand, seem to correlate with specific time reference either through temporals in the sentence or past (present perfect) marking on the verb. At least, universal readings are more difficult to recover under these premises. As a result of this coding procedure, we have now collected pairings of the three different kinds of *man* (universal, existential and ambiguous) and the corresponding English strategy. In the next section, I will turn to the actual quantification and analysis of the data.

8. Analysis

First, the data will be summed up and an overview will be provided. In the subsequent analysis, there will be focus on pronouns and strategies that imply a human agent. Given that passives and other non-pronominal strategies are so prominent in the data, we will have a look first at the distribution of pronouns vs. other strategies. In table 3 below, you can see that in both corpora there is almost a 50% ratio. This could pose a problem since the comparability of pronouns with other strategies like passives is disputable. However, if we assume that passives and to-infinitives imply human agents and treat them similarly, an observation whether there are systematic changes in the semantics is justified. In table 4, you can see that there are only few strategies that do not at least imply human agents. To make all human impersonal strategies comparable with each other, the notion of vagueness has been kept throughout this study. This treatment of passives and infinitives corresponds to the agent-defocussing view on impersonalization.

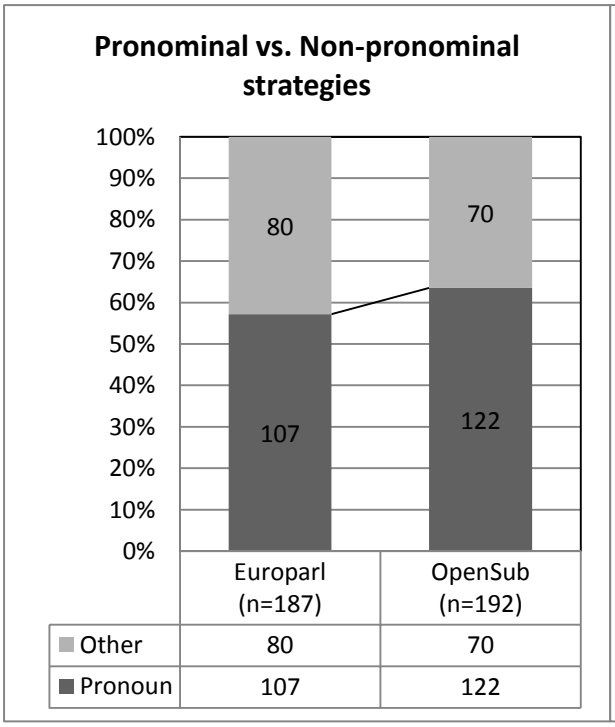


Figure 3: Pronominal vs. Non-Pronominal Strategies

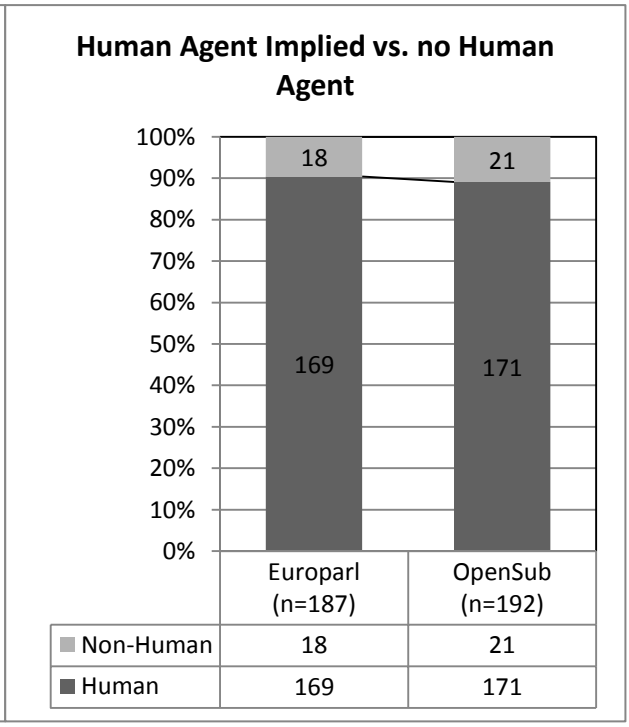


Figure 4: Human Agent Implied vs. no Human Agent

As is apparent from the figures, both corpora do not differ much in that respect. Although 44% of the translations in OpenSub contain *you*, the ratio of pronominal vs. non-pronominal is relatively equal. A chi-squared test results in a p-value of 0.2 (χ=1,583). The difference could still be incidental. That means that although *you* is overrepresented, other pronouns are rarer in turn. Nevertheless, *you* is the closest meaning equivalent in English and this suggests that meaning changes in OpenSub are less frequent. The challenge now is to quantify a variable for 'meaning change'.

8.1 From Translation Equivalent to Meaning Variable

As mentioned earlier, the equivalents of *man* function as indicator for meaning shifts, and so do the three readings of *man*. In order to capture differences between the English and the German sample, the strategies and the instances of German *man* have to be converted into nominal variables indicating their meaning. In table 1, you can see a summary of the features of each strategy as discussed in sections 3 and 5. A plus (+) indicates the

	Universal	Inclusion
man_univ	+	**±**
man_exist	-	-
you	+	**±**
one	+	**±**
we	+	+
they	-	-
people	-	**±**
I	+	+
indef	-	-
dem	**±**	**±**
passive	**±**	**±**
inf	**±**	**±**
NP_coll	+	**±**
NP_gen	+	**+**
NP_def	-	**-**
NP_other	n/a	n/a

Table 1:Overview of Strategy Features

feature in the table head, a minus (-) the counterpart[5] of it (existential and exclusive, respectively), and a plus-minus (±) stands for ambiguous.

Keep in mind that the passive and infinitive-constructions do not feature a pronoun or noun phrase. Since I do not assume zero pronouns, any information with respect to exclusivity or quantification of the referent is completely context-dependent. Consequently, all these strategies were coded as ambiguous for both parameters. Also note that the nominalizations other than collective nouns, generic humans and definite descriptions cannot properly be treated in the course of this paper. The other two strategies left over are adjectives and idioms. Together with nominalizations and other nouns they will form a group of 'others', which is equal to the group with non-human agents. In a next step all strategies receive a numeral (still nominal though) value, and meaning combinations across the German and English sentences irrespective of the particular strategies can be compared. Transferring the strategies into meaning variables has the big advantage that even extremely rare strategies can be sensibly treated because it is not the actual form of the expression, but its meaning that is focussed on.

8.2 Main Differences

The first thing investigated with the data is how many meaning differences there are all in all. The hypotheses H1 and H1a are tested. The ratios are given in figure 5. In this part of the analysis even the group 'others' contributes meaningfully because we treat no particular meaning change here. In the sample of the Europarl corpus only about 27% of the sentences are identical in meaning with respect to inclusion and quantification. For OpenSub the ratio is 46%. Both percentages are extremely high, considering these are sentences from

[5] Note that with *counterpart* I do not mean the logical opposite (non-universal).

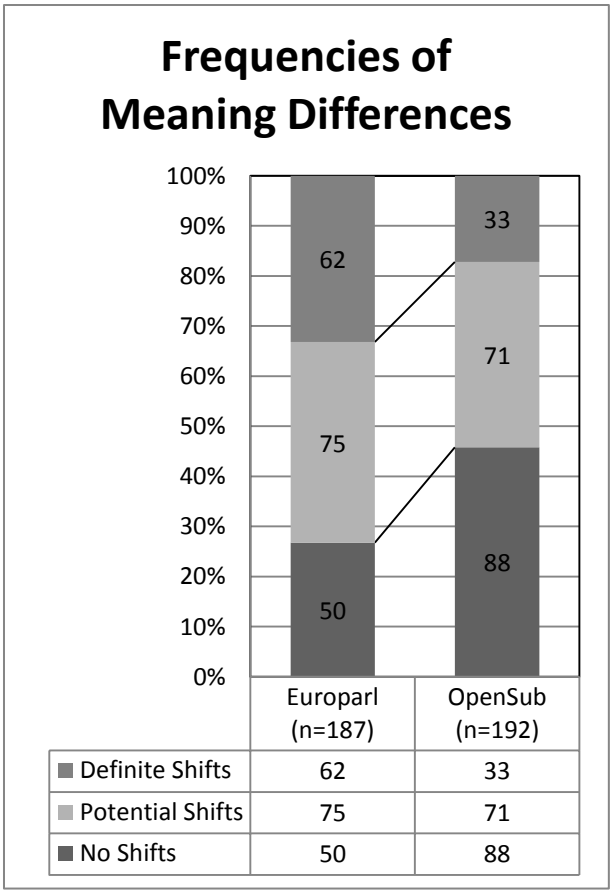


Figure 5: Frequencies of Meaning Differences

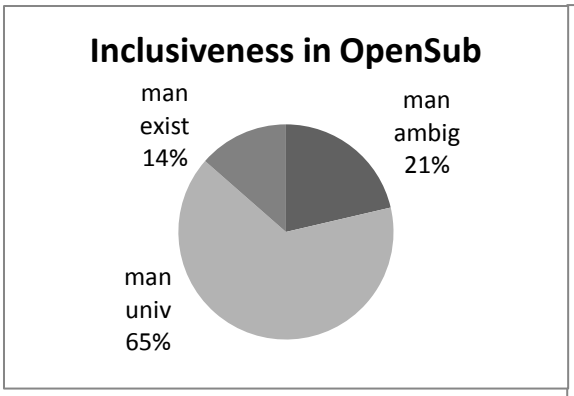
parallel corpora. However, since the context has been disregarded there is a huge bulk of sentence pairings that have at least potentially the same meaning. Those are the combinations which have an ambiguous feature vs. a definite or ambiguous one while the other parameter is constant. Keep in mind that these are not equal since remaining ambiguity can be the speaker's conscious choice. With that adjustment being done, there are still about 33% meaning shifts in Europarl and 17% in OpenSub. Whether or not H_1 can be supported depends on how ambiguous uses are evaluated. What is striking, however, is that there is a huge difference between both corpora. Although in Europarl the interpreters are motivated to stay as closely as possible to the original meaning, and although the texts are revised, the meanings cannot be mapped sufficiently. A possible explanation for this is that whenever the speakers in parliament start getting obscure and avoid addressing definite people, German speakers and interpreters have a convenient tool in the pronoun *man* whereas English speakers are forced to specify more due to their system of pronouns. In films, there is more leeway for creativity in translation, still the meanings map better although not sufficiently still. The chi-squared test for the values in figure 5 provides p-values far lower than 0.001.[6] The hypothesis H_1a can be strongly supported.

8.3 Quantification and Inclusion across English and German

If we look at the meaning components independently, we get the following picture for inclusion/exclusion. There are only few cases of existential *man* in both corpora.

[6] For 'no shifts' vs. rest: $p<0.001$; $\chi=14,919$
For potential shifts plus no shifts vs. rest: $p<0.001$; $\chi=12,859$

Figure 6: Inclusiveness in OpenSub



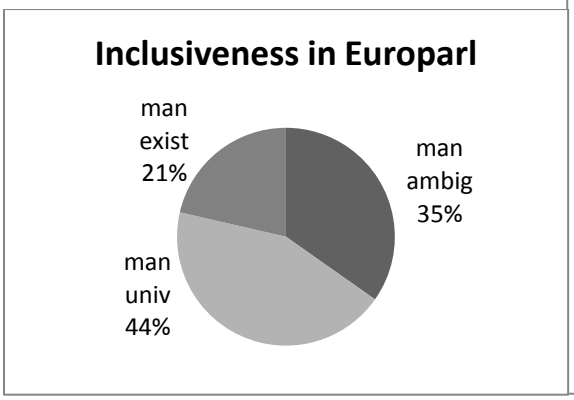


Figure 7: Inclusiveness Europarl

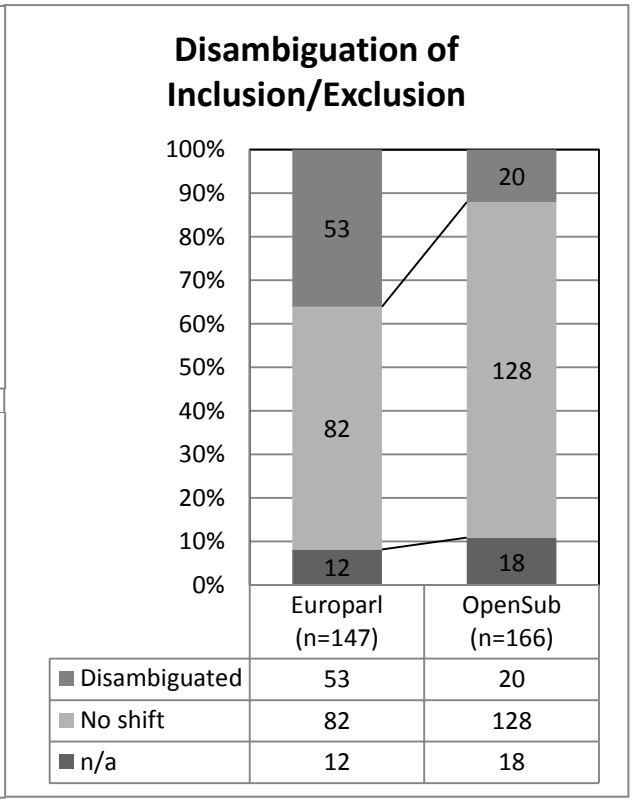


Figure 8: Disambiguation of Inclusion/Exclusion

That means that about 86% and 79%, respectively, are ambiguous with respect to inclusion/exclusion (cf. figure 5, 6). Taking a closer look at the equivalents of these sentences, one can see that they are not so often disambiguated in OpenSub but more so in Europarl. About 36% of the ambiguous and universal *man* sentences are disambiguated in Europarl which is 29% overall. Therefore, almost every third time there is a disambiguation of inclusive/exclusive in English relative to German. This is three times more frequent than in OpenSub (12% in the sub-sample, 10% overall). As a result, the hypothesis H_2 can be strongly supported for Europarl and less strongly, if at all, in OpenSub.

For the last hypothesis – English ambiguous strategies tend to be disambiguated in German –, the ambiguous strategies of English are the centre of analysis. If there were no discrepancies, demonstratives, passives and infinitives (cf. table 1) should be equivalent to ambiguous *man*. In Europarl, there are overall 54 occurrences of ambiguous *man* and in OpenSub there are 48. While 32 (59%) of these show either clearly universal or clearly existential features in Europarl, there are 39 out of 48 (81%) that do so in OpenSub. Considering the percentages, that seems a lot. However, the partial samples are too small to make a decision in favour of or against the hypothesis. Also the difference between the two

corpora, as bis as it may seem, could still be incidental. A chi-squared test receives a p-value of >0,1 and the phi-coefficient is about 2,24, i.e. the contingency is low, if not marginal, and the probability that it is by chance is also quite high. Although the other results were solid, the sample size has to be increased for further analyses on particular areas.

9. Critical Remarks and Conclusion

In this corpus study, I have shown that the English equivalents of *man* diverge in meaning. The pronominal strategies have considerable input themselves, whereas *man* has only little semantic substance. The non-pronominal strategies are problematic but statistically very prominent. The corpora Europarl and OpenSub do not only vary in the distribution of translation equivalents, but also in their accuracy. Nevertheless, these results have to be evaluated with the fact in mind that the contexts were mostly disregarded. Embedded into context, the equivalents in English could produce a higher accuracy in meaning. One could get into further detail taking in the sentence features veridicality and genericity. To get the full picture this would even be necessary since the meaning of the pronoun and the sentence meaning definitely interact. Meaning differences between both languages can be expected there, too. However, the coding of contexts is a lot more complicated and complex. Both ways would have to be interpreted and a lot more context than just one sentence would have to be looked at. This amount of coding effort exceeds the size and scope of this study. Corpus analysis is a sub-kind of textual analysis. This group of methods naturally involves a considerable input of the coder in form of interpretation. In this study, the input is minimized as much as possible. I tried to disambiguate as little as necessary, thereby reducing the distortion by my own intuition. For larger scale studies including context variables, there ought to be measurements of intra-coder reliability. Multiple coding of the same test samples in a pre-test will also enhance the criteria for coding variables such as universal/existential.

Taking the translation strategies themselves as indicators avoided much of the coder input. However, the analysis of the strategies is highly theory-dependent. Without a proper treatment of the non-pronominal strategies there is no exhaustivity in the variables, especially since they do not apply for the group of NPs, at all. There has to be further

treatment of non-pronominal strategies and their relationship to impersonals. There has been much effort on comparing pronominal strategies in the literature. There are some approaches to zero-subject constructions especially within the paradigm of Generative Grammar. However, the relation of passives and infinitives to pronominal impersonals remains unclear, though not being trivial as the data suggests. Without exhaustivity of the variables describing the impersonal strategies, the quality of any quantitative study is diminished. In Gast & van der Auwera (2013), in contrast to van der Auwera et al. (2012), the distinction of inclusion/exclusion is disregarded in favour of internal/external perspective. This distinction is motivated by Moltmann's (2010) concept of detached self-reference. The interpretation of impersonal pronouns is defined by the perspective suggested. Some impersonal pronouns offer an internal perspective, i.e. an instruction to imagine being someone else (Gast & van der Auwera 2013: 25). This could perhaps make better distinctions between *you* and *we*, for instance. This distinction is assumed not to be one that is encoded in the pronoun (Gast & van der Auwera 2013: 42). Therefore, it is context-dependent. Because I did not take in to account the context features, and because the coding of internal perspective is mostly a matter of intuition again, it was not taken up in this paper.

Some methodological problems remain also. The sample size is too small to allow generalization in most kinds of analyses. Generalization is a problem of corpus studies, anyway. One has to keep in mind that the results in this study are only relevant for the respective corpora. The relation to natural language is questionable. To get better results and a wider overview of translation strategies, the sample sizes have to be enlarged. Most strategies were just too rare to be analysed properly. The discard of sample sentences is also quite high. From overall 440 sentences there are 61 that are unusable. That means every seventh sentence cannot be used in this design. Especially the status of non-translations probably has to be taken into account in a quantitative study. The method is also arguably lopsided in that only *man* is analysed on the German side. To get the full picture of cross-linguistic variation between English and German impersonalization strategies, the other German strategies have to be quantified, as well. This is, however, hardly possible as discussed above. There is one possibility for large-scale quantitative studies of impersonals that take into account more impersonal strategies on both sides. POS (part-of-speech) tagging could help identifying certain constructions. The corpora used in this paper are not

tagged. However, it is highly unlikely to find or create corpora with tagging for impersonal uses of pronouns, with the concept of impersonalization being so fuzzy. Nevertheless, for the scope of this paper, the design was sufficient and the results offer an overview of English strategies associated with German *man*. The meaning of impersonal strategies in parallel corpora seems to diverge. This has implication, for example, for other methodical designs that assume meaning equivalency across languages in parallel corpora.

List of Abbreviations

ACC	accusative
adj	adjective
anaph	anaphora
DAT	dative
deic	deictic
dem	demonstrative pronoun
Europarl	European Parliament Proceedings Parallel Corpus
exist	existential
HIP	human impersonal pronoun
indef	indefinite pronoun
inf	infinitive
NP	noun phrase
NP_coll	collective noun
NP_def	definite description
NP_gen	generic human
non_imp	non-impersonal
OpenSub	Open Subtitles Corpus
part_past	past participle
part_pres	present participle
POS	part-of-speech
psv	passive
univ	universal

Corpora

Europarl – European Parliament Proceedings. Available online at <http://opus.lingfil.uu.se/Europarl3.php>. 29 July 2013.

OpenSubtitles. <http://www.opensubtitles.org>*.* Available online at <http://opus.lingfil.uu.se/OpenSubtitles.php>. 29 July 2013.

OPUS – the open parallel corpus. Available online at <http://opus.lingfil.uu.se/index.php>. 29 July 2013.

Bibliography

van der Auwera, Johan, Volker Gast, and Jeroen Vanderbiesen (2012). "Human impersonal pronouns in English, Dutch and German". *Leuvense Bijdragen* 98:1. 24-64. Available online at <http://www.personal.uni-jena.de/~mu65qev/papdf/oldenburg_pp.pdf>.

Bhatt, Rajesh & Roumyana Izvorski (1997). "Genericity, Implicit Arguments and Control". *Proceedings of SCIL, MITWPL*.

Bhatt, Rajesh & Roumyana Pancheva (2005). "Implicit Arguments". In: *The Blackwell Companion to Syntax*. eds. M. Everaert and H. van Riemsdijk. Washington: Blackwell. Available online at <http://www.academia.edu/download/30910732/bhatt-pancheva-imp.pdf>. 23 August 2013.

Blevins, James P. (2003). "Passives and impersonals." *Journal of Linguistics* 39: 473-520.

Cabredo Hofherr, Patricia (2003). "Arbitrary readings of 3PL pronominals". In: *Proceedings of the conference 'sub-7 – Sinn und Bedeutung',* Konstanz. Ed. Matthias Weisgerber. Available online at <http://ling.unikonstanz.de/pages/conferences/sub7/proceedings/download/sub7_hofherr.pdf>. 9 August 2013.

Cabredo Hofherr, Patricia (2006). "'Arbitrary pro and the theory of pro-drop". In: *Agreement and arguments*. Eds. P. Ackema, M. Schoorlemmer and F. Weerman. Oxford: Oxford University Press. 230-258.

Chomsky, Noam (1995). *The minimalist program*. Cambridge, Mass.: MIT Press.

Dimowa, Anna (1981). "Die Polysemie des Pronomens *man* in der deutschen Gegenwartssprache und die Kontextbedingungen für seine Monosemierung". *Beiträge zur Erforschung der deutschen Sprache* 1: 47-75.

Duden (2013). “man“. <www.duden.de/rechtschreibung/man_jemand_irgendeiner_irgendeine>. 21 August 2013.

Gast, Volker & Johan van der Auwera (2013). “Towards a distributional typology of human impersonal pronouns, based on data from European languages.” In: *Languages across boundaries.* eds. Bakker, D. & M. Haspelmath. Available online at <http://www.personal.uni-jena.de/~mu65qev/papdf/anna_pp.pdf>. 19 August 2013.

Giacalone Ramat, Anna & Andrea Sansò (2007). “The spread and decline of indefinite man-constructions in European Languages: An Areal Perspective“. In: *Europe and the Mediterranean as linguistic areas: Convergences from a historical and typological perspective*. Ed. P. Ramat and E. Roma. 95-131. Amsterdam: Benjamins.

Kitagawa, Chisato & Adrienne Lehrer (1990). “Impersonal uses of personal pronouns”. *Journal of pragmatics* 14(5): 739-759.

Koehn, Philipp (2005). “Europarl: A parallel corpus for statistical machine translation”. *MT Summit 2005*. Available online at <http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf>. 21 August 2013.

Kratzer, Angelika (2000). “Generic Pronouns and Logophoricity“. Available online at <http://www.academia.edu/2857154/German_impersonal_pronouns_and_logophoricity>. 8 August 2013.

Langacker, Ronald W. (1991). *Concept, image, symbol: The cognitive basis of grammar*. Berlin: Mouton de Gruyter.

Langenscheidts Großwörterbuch der englischen und deutschen Sprache (2004). Ed. H. Willmann. Berlin: Langenscheidt.

Lehmann, Christian (1995). *Thoughts on grammaticalization*. Munich: Lincom.

Moltmann, Friederike (2010). “Generic *one*, arbitrary PRO, and the first person”. *Natural Language Semantics* 14: 257-281.

Siewierska, Anna (2008). “Introduction: Impersonalization from a subject-centred vs. agent-centred perspective”. *Transactions of the Philological Society* 106: 115-137.

Siewierska, Anna (2011). “Overlap and complementarity in reference impersonals: Man-constructions vs. third person plural-impersonals in the languages of Europe”. In: *Impersonal constructions: A cross-linguistic perspective*. eds. A. Siewierska and A. Malchukov. Amsterdam: Benjamins. 57-89.

Siewierska, Anna & Maria Papastathi (2011). “Third person plurals in the language of

Europe: Typological and methodological issues". *Linguistics* 43: 575-610.

Tiedemann, Jörg (2009). "News from OPUS – A collection of multilingual corpora with tools and interfaces". In: *Recent advances in natural language processing* 5. eds. N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov. Amsterdam: Benjamins. 237-248. Available online at <http://stp.lingfil.uu.se/~joerg/published/ranlp-V.pdf>. 21 August 2013.

Wales, Katie (1996). *Personal pronouns in present-day English*. Cambridge: Cambridge University Press.

Zifonun, Gisela (2001). "Man lebt nur einmal. Morphosyntax und Semantik des Pronomens *man*". *Deutsche Sprache* 3: 232-253.

Zwarts, Frans (1995). "Nonveridical contexts". *Linguistic Analysis* 25.3/4: 286-312.

Declaration of Academic Integrity

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbst angefertigt und alle von mir benutzten Hilfsmittel und Quellen angegeben habe; alle wörtlichen Zitate und Entlehnungen sind als solche gekennzeichnet.

27.08.2013	
Datum	Unterschrift