A Corpus-based Study of the German Recipient Passive

Patrick Ziering, Sina Zarrieß, Jonas Kuhn

Institute for Natural Language Processing
University of Stuttgart
[Patrick.Ziering, Sina.Zarriess, Jonas.Kuhn] @ims.uni-stuttgart.de

Abstract

In this paper, we investigate the usage of a non-canonical German passive alternation for ditransitive verbs, the *recipient passive*, in naturally occuring corpus data. We propose a classifier that predicts the voice of a ditransitive verb based on the contextually determined properties of its arguments. As the recipient passive is a low frequent phenomenon, we first create a special data set focussing on German ditransitive verbs which are frequently used in the recipient passive. We use a broad-coverage grammar-based parser, the German LFG parser, to automatically annotate our data set for the morpho-syntactic properties of the involved predicate arguments. We train a Maximum Entropy classifier on the automatically annotated sentences and achieve an accuracy of 98.05%, clearly outperforming the baseline that always predicts active voice (94.6%).

1. Introduction

Besides the regular passive voice, many languages exhibit so-called non-canonical passive constructions as for instance the *get* passive in English, the *se faire* passive in French or the recipient passive (also called *bekommen* passive) in German which will be the focus of this work. The genesis of these non-canonical passives and their grammatical restrictions have been subject to extensive debates in theoretical linguistics. In this work, we are interested in the usage of this alternation which we consider as a meaning-equivalent paraphrase of the regular passive and the active. This is illustrated in Example (1) which presents the realisation of a German ditransitive verb in active, passive and recipient passive voice.

(1) a.   Der Kellner serviert den Gästen eine Flasche
          The waiter serves   the guests a   bottle
          Wein.
          wine.
          "The waiter serves the guests a bottle of wine."
     b.   Eine Flasche Wein wird den Gästen vom Kellner
          A   bottle wine is   the guests by the
          serviert.
          waiter served.
          "A bottle of wine is served to the guests by the waiter."
     c.   Die Gäste bekommen vom Kellner eine
          The guests get   by the   waiter
          Flasche Wein serviert.
          a   bottle wine   served.
          "The guests get served a bottle of wine by the waiter."

The aim of our study is to be able to predict the voice of a German ditransitive verb from its given context in a corpus sentence. More precisely, we pursue the hypothesis that the German recipient passive can be predicted from the contextually determined properties of its arguments (i.e. the definiteness of the agent, the person of the patient etc.). This hypothesis follows the work by Bresnan et al. (2005) who show that the usage of the English dative alternation is guided by statistical tendencies which can be modeled on the basis of multiple contextual factors.

Since the corpus-based modelling of linguistic alternations typically requires relatively "deep" linguistic information (e.g. morphological information about definiteness etc.), previous approaches usually rely on manually annotated treebank data. However, in the case of the German recipient passive, the construction is too low frequent in order to be appropriately represented in standard German resources. Therefore, in a first step, we built a specialised data set of occurrences of German ditransitive verbs, extracted from a set of German corpora. This data set was automatically annotated with the German LFG parser (Rohrer and Forst, 2006). In a second step, we exploit the automatically annotated sentences to train and test a maximum entropy classifier that is given the context of a ditransitive verb and predicts its voice. The classifier achieves an accuracy of 98.05% over a baseline of 94.6% which always predicts the active voice.

While previous work on the German recipient passive has mostly dealt with syntactic, semantic and stylistic restrictions on the phenomenon (Diedrichsen, 2004), the results presented in this paper suggest that the usage of the German recipient passive alternation is subject to soft statistical constraints that can be modelled by means of multiple contextual factors. The constraints we observed are very similar to factors found in previous work on different alternations and languages.

The remainder of the paper is structured as follows. We provide some background on related work in Section 2., and on the linguistic phenomenon in Section 3. In Section 4., we describe the construction of our data set. Our classification experiments and their results are presented in Section 5. In Section 6, we look at some of the features in more detail in order to see whether the model conforms to linguistic tendencies observed in previous works.

2. Related Work

There are a number of (mostly English) argument alternations or syntactic paraphrases that have been investigated in corpus-based linguistic studies, such as the English passive (Thompson, 1987; Bresnan et al., 2001), the English dative alternation (Bresnan et al., 2005; Arnold et al., 2000) or *that*-clause reduction (Levy and Jaeger, 2007). In these works, the use of an argument alternation in language is typically related to a number of properties of their arguments, e.g. definiteness, animacy or complexity of the involved noun phrases. In Aissen (1999) and Aissen (2003), this idea is described in terms of markedness hierarchies (which can be formalised as Optimality-theoretic constraints). These hierarchies associate argument functions with certain prototypical properties for person, animacy, or definiteness; e.g., subjects tend to be higher on the person scale than objects (1st person > 3rd person).

Depending on the language, these markedness hierarchies can be encoded as hard constraints in the grammar, or figure as soft, usage-related preferences. For the passive alternation, this contrast is nicely established in Bresnan et al. (2001) who investigate the relationship between voice and person in English and Lummi. In Lummi, the person hierarchy is encoded as a hard constraint into the grammar of the language such that the subject must be higher in the person hierarchy than the object. This hierarchy is illustrated in (1), meaning that in the case of a third-person agent and a first person patient, the verb is obligatorily passivised in order to be able to realise the first person entity as the subject.

$$1.pers \cup 2.pers > 3.pers \qquad (1)$$

Obviously, in English, there is no such hard constraint, both sentences in Example (2) are grammatical. Nevertheless, Bresnan et al. (2001) found a clear statistical tendency for English passivation in cases where the Lummi hard constraint would be violated, indicating that this hard constraint is present as soft constraint in English. In our work, we hope to find similar factors accounting for preferences of the recipient passive over the regular passive and the active.

(2) a. The man$_{3ps}$ sees us$_{1ps}$
    b. We$_{1ps}$ are seen by the man$_{3ps}$

In a similar setting, Bresnan et al. (2005) investigate the English dative alternation. Verbs like *give* can have two variants of complements: two objects (the *double object construction*) or an object and a prepositional phrase, see Example (3).

(3) a. Susan gives the children toys
    b. Susan gives toys to the children

A possible explanation for this syntactic variation could be the meaning of the event: is it a kind of movement (use of a prepositional phrase) or a change of state (use of a double object construction)? Actually, similar, purely meaning-based explanations have been proposed for the German recipient passive, as we discuss in the following Section.

However, Bresnan et al. (2005) show that the alternation can be modeled taking into account a number of contextually determined argument properties, such as the complexity of the noun phrases, definiteness, pronominality, salience, number, and person. Based on these features, Bresnan et al. (2005) fit a logistic regression model that yields 94% prediction accuracy.

The methodology that is typically pursued in these theoretically motivated works on argument alternations is to hand-code a set of selected sentences where a specific alternation is displayed. By this means, the authors isolate the phenomenon under investigation relying on high-quality, manually annotated data. As a result, the obtained models allow for a detailed analysis of the individual contextual factors at play.

A somewhat different approach is taken in recent works from the domain of Natural Language Generation showing the relevance of syntactic alternations for computational applications. Rajkumar and White (2011) model the realisation of *that*-clauses in a grammar-base generation component taking into account the features proposed in Levy and Jaeger (2007). Zarrieß et al. (2011) generate passives interacting with free word order in German based on a linguistically-informed feature model that incorporates the relative hierarchy of the involved arguments.

In contrast to the previously mentioned theoretical works on argument alternations, these application-oriented works typically deal with more data where several syntactic variation phenomena interact. Also, Zarrieß et al. (2011) conduct their experiments on automatically parsed data, meaning that the statistical component has to deal with noise introduced from parsing errors. A further difference to investigations based on hand-coded data is the design of the feature model, and the type of argument properties taken into account. While Bresnan et al. (2005) encode the information status (or givenness) of the predicate arguments, the feature models in Rajkumar and White (2011) and Zarrieß et al. (2011) are based on morpho-syntactic information that can be extracted from automatically produced syntactic analyses. However, the information status of a noun phrase is often reflected in its morpho-syntactic realisation, such that these automatically obtainable features typically provide relatively accurate models (Zarrieß et al., To appear).

The study of the German Recipient Passive that we present in this paper is clearly inspired from the theoretical approach pursued in (Bresnan et al., 2001) or (Aissen, 1999). However, our ultimate goal is to inform generation-based approaches such as Zarrieß et al. (2011), where deep linguistic features such as informations status cannot be assumed to be available.

3. The Recipient Passive

In German, certain ditransitive verbs can be passivised in two different ways: (i) the patient or direct object can be promoted to a subject, (ii) the recipient or dative object can be promoted to a subject while the agent is demoted to an optional prepositional object. Both passives, the regular passive (i) and the recipient passive (ii) are formed by means of an auxiliary and the past participle of the main

verb. The auxiliary of the regular passive is *werden* (in English: *be*), while in the recipient passive *bekommen* (*get*) has to be used. Note that the verb *bekommen* (*get*) can also be used as a full verb in German.

The two main questions debated in the linguistic literature are (i) which class of ditransitive verbs or dative objects licenses the recipient passive?, (ii) what is the linguistic origin of the construction and to what extent is the construction with the verb *bekommen* fully grammaticalised?

According to the theory of grammaticalisation (Diewald, 1997), the recipient passive originates from a construction where *bekommen* is used as a main verb (i.e. *get*) and its object is modified by a predicative participle (cf. (4-a)). In these constructions, the main verb started to lose its original meaning such that usages as in Example (4-b) became possible, where *bekommen* has two possible readings: the main verb and the auxiliary (cf. (4-b)). In the final, fully grammaticalised stage, the main verb reading is not available and the only possible reading of *bekommen* is the auxiliary, i.e. the recipient passive (cf. (4-c)).

(4) a. Sie bekommt die Bretter vom   Schreiner
        She gets   the planks from the carpenter
        schon passend   zugeschnitten.
        already appropriately cut.
        "The carpenter gives her the planks which are already cut to the appropriate size."
    b. Sie bekommt den Katalog   zugeschickt.
        She gets   the catalogue sent.
        "She gets the catalogue sent by post" or "The catalogue is sent to her by post"
    c. Peter bekam den Führerschein weggenommen
        Peter gets   the licences   taken away.
        "Peter's driver licence was suspended."

The German recipient passive occurs most frequently with the auxiliary *bekommen*, which is stylistically relatively unrestricted. Pimanyonok (2004) mentions two further auxiliaries, that can only be used in specific contexts and textual genres: *erhalten* (*obtain*) and *kriegen* (*cop*). The auxiliary *erhalten* is used only in instructions and non-fictional texts, whereas *kriegen* is very infrequent in written language and is only used colloquially. The examples in (5) show the three possible versions of the recipient passive.

(5) a. Er bekam$_{universal}$ ein Buch geschenkt.
        He got   a   book given.
        "He was given a book."
    b. Sie erhält$_{literary}$ den Nobelpreis verliehen.
        She got   the Nobel Prize awarded.
        "She was awarded the Nobel Prize."
    c. Die Frau   kriegt$_{colloquial}$ die Bretter
        The woman gets   the planks
        zugeschnitten.
        cut.
        "The woman is given the planks cut to the appropriate size."

Grammatically, not stylistically, *bekommen* and *kriegen* are equivalent. They can both be used to passivise ditransitive verb having the meaning of "removing". This is not possi-

ble with the auxiliary *erhalten*, see Example (6).

(6) a. * Peter erhält den Führerschein   entzogen.
        * Peter gets   the driver's license revoked.
        "Peter's driver's license is revoked."
    b. Peter bekommt den Führerschein   entzogen.
        Peter gets   the driver's license revoked.
        "Peter's driver's license is revoked."

In opposition to the regular passive (cf.   (7-a)), an impersonal passive (i.e. a construction without subject) is not possible in the recipient passive (cf. (7-b)) (Pittner and Berman, 2004):

(7) a. Heute wird getanzt.
        Today is   danced.
        "Today, people dance. "
    b. * Heute bekam verziehen
        * Today, was   forgiven.
        "Today, people forgave."

There are several restrictions on the main verb that can undergo the recipient passive alternation. As described in (Diedrichsen, 2004), ditransitives that do not subcategorise a recipient cannot undergo the alternation, see Example (8).

(8) a. Du schuldest mir 100 Euro.
        You owe   me 100 Euro.
    b. *Ich bekomme von dir 100 Euro geschuldet.
        I   get   from you 100 Euro owed.

In some cases, the main verb may only have a two place valence as in (9).

(9) Er bekommt verziehen.
    He gets   forgiven.
    "He is forgiven."

The dative object in the active voice of a main verb must not be a reflexive pronoun (cf. (10)).

(10) a. Er stellt sich   die Frage   ...
         He puts himself the question ...
         He is wondering whether . . .
     b. [*]Er bekommt von sich   die Frage
         He gets   by himself the question
         gestellt . . .
         put   . . .

In the literature, the syntactic status of the recipient passive is highly controversial. One reason for this controversy is the remaining semantics of the auxiliary. Thereby, constructions with verbs of *removing* (the opposite of the auxiliary's meaning) seem ungrammatical, see Example (11). Consider, however, the counter-example .

(11) ?Er bekam das Fahrrad gestohlen.
     He got   the bike   stolen.
     "His bike was stolen from him."

Moreover, the morpho-syntactic ambiguity of *bekommen* may lead to ambiguous interpretations of the passivised events. The sentence in (12) has three readings.

(12) Peter bekommt einen Apfel gewaschen.
    Peter gets   an   apple washed.
    "Someone washes an apple for Peter." or:
    "Peter succeeds in washing an apple." or:
    "Peter is given a washed apple."

Another difference between the regular passive and the recipient passive is the importance of the prepositional agent for the information structure and the accent of the sentence. In (13-a) the prepositional agent is the most important (focused) element following the verb, whereas in (13-b) it is focused together with main verb, that constitutes the most important element. This difference gets apparent in the sentence accent (Eisenberg, 1999)

(13) a. Der Karl bekommt von der Karla das Formular
         The Karl gets   by the Karla the form
         ausgefüllt.
         filled in.
         "The form is filled in for Karl by Karla."
     b. Das Formular wird dem   Karl von der Karla
         The form   is   to the Karl by the Karla
         ausgefüllt.
         filled in.
         "The form is filled in for Karl by Karla."

The recipient passive is often used in a coordination reduction (i.e. a coordination of sentences with coreferent subjects where only the first one is mentioned) as in Example (14-a). If a speaker wanted to realise Example (14-a) without a recipient passive, he would have to coordinate two sentences with different subjects which is information-structurally suboptimal (cf. Example (14-b)) (Diedrichsen, 2004).

(14) a. Er fuhr zu schnell, wurde von der Polizei
         He drove too fast,   was by the police
         angehalten und bekam den Führerschein
         stopped   and get   the drivers license
         entzogen
         withdrawn.
         "He drove too fast, was stopped by the police
         and lost his driver's license."
     b. Er fuhr zu schnell und die Polizei entzog
         He drove too fast   and the police took away
         ihm   den Führerschein.
         from him the drivers license.
         "He drove too fast and the police confiscated his
         driver's license."

In this work, we will leave aside more details on this discussion of lexical and grammatical constraints of the recipient passive. Instead, we are interested in contexts where a ditransitive verb potentially licenses all three voices (active, regular passive and recipient passive), but the speaker chose to realise the latter one. We will only consider the most frequent auxiliary *bekommen* for our study in order to exclude the factor of stylistic preference as much as possible.

4. Building the Data Set

Since the German recipient passive is constructed with a special auxiliary (*bekommen*), it can be reliably extracted from corpora by means of flat PoS patterns. We extracted all sentences with an occurence of the recipient passive from several German text collections. We considered the following corpora: (i) the **Huge German Corpus (HGC)**, a collection of newspaper texts, (iii) **DEWAC**, the German web corpus, (iv) the **GUTENBERG corpus**, that contains classic German literature of more than 550 authors, and (v) the **TIGER corpus**, a syntactically annotated treebank. For the extraction, we used the Corpus Query Processor (CQP), a component of the IMS Corpus Workbench (CWB) (Christ, 1994), that provides serveral German PoS-tagged corpora. The main advantage of this tool is the efficient specification and execution of queries as regular expressions within a given context.

corpus (size in sentences)	recipient passives	
	total freq.	rel. freq
HGC (∼ 12 Mio.)	4.664	0.04%
DEWAC (∼ 33 Mio.)	23.235	0.08%
GUTENBERG (∼ 6.5 Mio.)	966	0.02%
TIGER (36.475)	9	0.025%

Table 1: Occurences of recipient passives in German corpora

Table 1 displays the number of instances we found, showing that the recipient passive is used rarely with respect to the entire set of sentences in a corpus. However, the total number of recipient passive instances we found in our set of corpora is actually largely sufficient for statistical studies on the phenomenon.

Moreover, we note that there is a small set of main verbs in this set of sentences which seems to frequently realise the recipient passive. In Table 2, we show the main verbs ordered by their absolute frequency in the extracted recipient passive sentences.

Recall that in our usage-based study, our aim is to investigate the recipient passive as a syntactic paraphrase of the active and the passive voice. Therefore, we have to include all active and passive realisations of the respective verbs in our data set. However, including all verbs that occur at least once with a recipient passive in our data set would result in a sentence collection where the recipient passive corresponds to a very small portion of the realised voices. For instance, the verb *sagen* (*say*) occurs in 1,013,324 sentences, whereas only 529 sentences are in recipient passive (≈ 0.052%).

As a solution, we compute the relative frequencies of the recipient passive for each of listed verbs, and limit our data set to those verbs that have a high proportion of realisations in this voice. In Table 3, we present the top 11 list of verbs that have the highest proportion of recipient passive realisations. We note that this subset of verbs is quite coherent with respect to the broad semantic field, i.e. all verbs can be caracterized as "'*transfer*" verbs (that is the converse of the main verb's meaning of the auxiliary *bekommen*.).

Adopting the perspective of the recipient passive as a para-

verb	# occurences
schenken (*give*)	2,129
stellen (*supply*)	829
bieten (*offer*)	826
anbieten (*provide*)	698
zuweisen (*assign*)	653
verleihen (*award*)	622
vermitteln (*broker*)	618
zusprechen (*grant*)	568
zuschicken (*send*)	545
sagen (*say*)	529
erstatten (*refund*)	513
liefern (*deliver*)	513
zuteilen (*assign*)	445
servieren (*serve*)	426
bezahlen (*pay*)	419
überreichen (*hand over*)	408
verpassen (*inflict*)	394
vorsetzen (*present with*)	377
präsentieren (*present*)	376
zeigen (*show*)	344
(in die Hand) drücken (*thrust (into hand)*)	323
auszahlen (*disburse*)	316
anzeigen (*display*)	312
aushändigen (*surrender*)	289

Table 2: Absolute frequency of ditransitives in the recipient passive (2)

verb	# occurences	rec. pass.
zuschicken (*send*)	2,854	19.1%
vorsetzen (*present with*)	2,247	16.8%
zuteilen (*assign*)	3,789	11.7%
zusprechen (*grant*)	5,925	9.6%
aushändigen (*deliver*)	3,286	8.8%
zuweisen (*allocate*)	7,831	8.3%
schenken (*give*)	40,371	5.3%
servieren (*serve*)	8.373	5,1%
erstatten (*refund*)	10.200	5,0%
überreichen (*hand over*)	9.616	4,2%
verleihen (*award*)	33.162	1,9%

Table 3: Frequency of the recipient passive for specific ditransitives

phrase, we also need to consider the active and regular passive instances of these verbs in order to be able to compare their contexts. Thus, we extracted all instances of our 11 most frequent verb lemmas from our total set of corpora. The resulting data is a collection of 127,654 sentences. In Table 3, we report the relative frequency of the recipient passive for this specialised set of ditransitive sentences. While the occurence of a recipient passive can be detected

using flat PoS patterns, we need deeper linguistic information in order to appropriately analyse the respective contexts of the construction. We use a hand-crafted, broad-coverage Lexical Functional Grammar (Bresnan, 2001) for German (Rohrer and Forst, 2006) for preparing our data set. It integrates a statistical disambiguation module and achieves parsing coverage of about 80% on German newspaper text. The syntactic analyses are encoded on two representation levels: constituent structure (c-structure) which encodes constituency and linear order and functional structure (f-structure) representing grammatical functions and morphosyntactic features.

(15) a. Er bekommt von seiner Mutter ein Auto
He gets by his mother a car
geschenkt.
given.
"He has been given a car by his mother."

As an example, Figure 1 displays the f-structure for the sentence (15) derived by the German LFG. The f-structure contains a range of syntactic and morphological features that will be very useful for our classification experiments in Section 5., e.g. number and person for nouns (NUM and PERS), nominal and pronominal types (PRON-TYPE and NTYPE), definiteness (DET-TYPE).

5. Classification Experiments

In this section, we present a classifier that predicts the voice of a ditransitive verb, given its intrasentential context. We used MegaM, a maxium entropy classifier to build a log-linear model for voice prediction.

Since f-structures are syntactic representations which encode the voice of the given ditransitive verb, we have to hide this information to the classifier. We used a small set of hand-written rules in order to map the grammatical functions onto semantic roles. Thus, the subject of a passive is mapped to a theme or recipient, whereas the subject of an active is mapped to an agent.

Furthermore, we have to consider that the realisation of an agent is optional in both regular and recipient passive. If the ditransitive verb does not realise an agent, the probability of the active voice will be zero. To eliminate these trivial cases from our classification, we divide the data into two subsets: (i) ditransitives which realise all 3 roles (agent, recipient, theme), (ii) ditransitives which realise 2 roles (recipient, theme). In Table 4, we report the distribution of the voices in the two subsets.

	active	passive	recipient passive
agent	94.6%	3.3%	2.1%
no agent	0.0%	61%	39%

Table 4: Voice distribution for ditransitive realisations

Our log-linear model for voice prediction is based on the following types of features: (i) argument type (noun, proper name, pronoun), (ii) person, number and gender of the argument, (iii) syntactic complexity of the argument (coordinated, modified by relative clause, etc.), (iv) surface order of the arguments, (v) lemma of the main verb. The set of

```
            "Er bekommt von seiner Mutter ein Auto geschenkt"

       ⎡PRED          'schenken<[99:von], [258:Auto], [4:pro]>'                      ⎤
       ⎢            4⎡PRED  'pro'                                                  ⎤  ⎥
       ⎢SUBJ      655⎢NTYPE [NSYN pronoun]                                         ⎥  ⎥
       ⎢          658⎢                                                             ⎥  ⎥
       ⎢          740⎣CASE nom, GEND masc, NUM sg, PERS 3, PRON-FORM sie, PRON-TYPE pers⎦⎥
       ⎢            ⎡PRED       'von<[128:Mutter]>'                              ⎤  ⎥
       ⎢            ⎢           ⎡PRED  'Mutter'                                 ⎤ ⎥  ⎥
       ⎢            ⎢           ⎢CHECK ⎡_SPEC-TYPE [_COUNT +, _DEF +, _DET attr]⎤⎥ ⎥  ⎥
       ⎢            ⎢           ⎢      ⎣_INFL      strong-det                   ⎦⎥ ⎥  ⎥
       ⎢OBL-AG   223⎢OBJ 1206   ⎢NTYPE ⎡NSEM [COMMON count]⎤                     ⎥ ⎥  ⎥
       ⎢        2122⎢           ⎢      ⎣NSYN common        ⎦                     ⎥ ⎥  ⎥
       ⎢         128⎢           ⎢      ⎡     ⎡PRED 'pro'                       ⎤⎤⎥ ⎥  ⎥
       ⎢      99 1054⎢          ⎢SPEC  ⎢POSS ⎣GEND masc, NUM sg, PERS 3, PRON-FORM ihre⎦⎥⎥ ⎥  ⎥
       ⎢     929 3286⎢          ⎢                                               ⎥ ⎥  ⎥
       ⎢    2711 2206⎢          ⎣CASE dat, GEND fem, NUM sg, PERS 3              ⎦ ⎥  ⎥
       ⎢    2715    ⎣PSEM dir, PTYPE sem                                         ⎦  ⎥
       ⎢            ⎡PRED  'Auto'                                              ⎤    ⎥
       ⎢            ⎢CHECK [_SPEC-TYPE [_COUNT +, _DET attr]]                  ⎥    ⎥
       ⎢         318⎢NTYPE ⎡NSEM [COMMON count]⎤                               ⎥    ⎥
       ⎢OBJ     1340⎢      ⎣NSYN common        ⎦                               ⎥    ⎥
       ⎢        1738⎢      ⎡   ⎡PRED     'eine'  ⎤⎤                            ⎥    ⎥
       ⎢         258⎢SPEC  ⎣DET⎣DET-TYPE indef   ⎦⎦                            ⎥    ⎥
       ⎢        1323⎢                                                          ⎥    ⎥
       ⎢    350 1824⎢                                                          ⎥    ⎥
       ⎢   1544 1793⎣CASE acc, GEND neut, NUM sg, PERS 3                       ⎦    ⎥
       ⎢   1547     ⎡_AUX-FORM <bekommen-pass>                ⎤                     ⎥
       ⎢   2380 CHECK⎢_VLEX    [_AUX-SELECT haben]             ⎥                     ⎥
       ⎢   2427     ⎢                                         ⎥                     ⎥
       ⎢     69     ⎣_VMORPH  [_PARTICIPLE perfect]           ⎦                     ⎥
       ⎢   3397                                                                     ⎥
       ⎢   2255 TNS-ASP   [MOOD indicative, PASS-ASP dynamic_, TENSE pres]          ⎥
       ⎢   2286 TOPIC     [4:pro]                                                   ⎥
       ⎣   2290 CLAUSE-TYPE decl, PASSIVE bekommen, VTYPE main                      ⎦
```
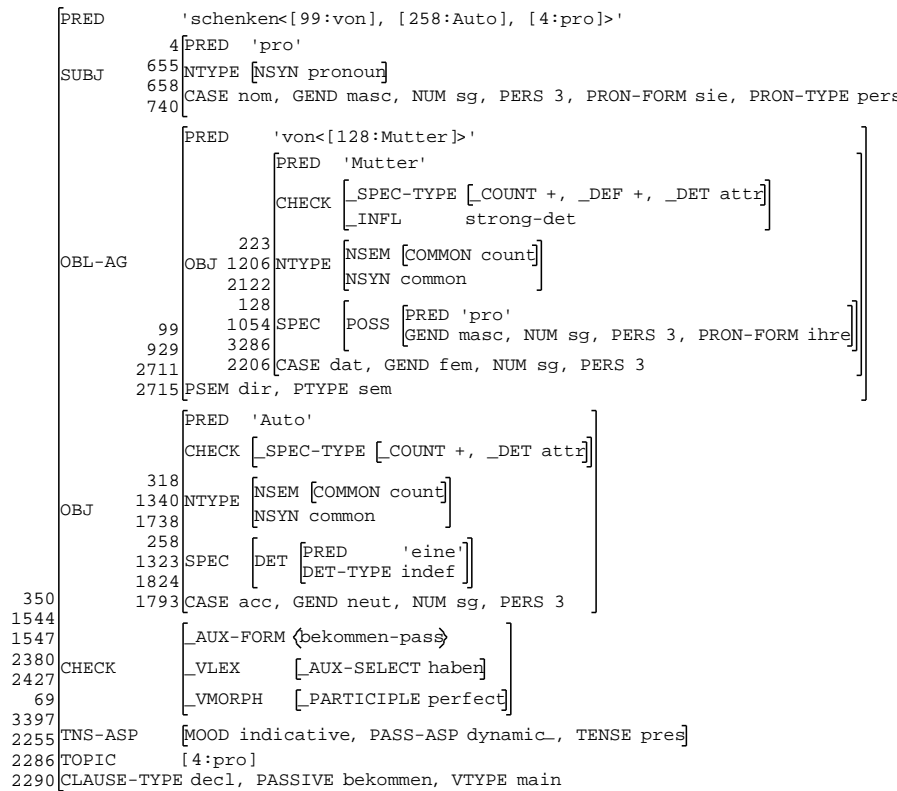
Figure 1: F-structure for sentence (15)

features is listed in Table 8. For each feature, we also report its frequency for a given argument role in its different voice realisations. This frequencies will be further discussed in Section 6. Note that, in contrast to Bresnan et al. (2001), our features do not include a direct encoding of the informations status or "newness" of a given argument since these cannot be extracted from an LFG f-structure. We hypothesise that we can approximate these properties to a certain extent via morpho-syntactic properties of the arguments.

We randomly divide our data set into 90% training sentences and 10% test sentences for both subsets (agent and no agent). In the agent subset, the classifier achieves an accuracy of 98.05% significantly beating the baseline which always predicts an active voice. Since this data set is so heavily unbalanced, it is also interesting to look at the precision and recall for the different voices (Table 5). Despite the strong balance towards the active, the prediction of the recipient passive is still precise with a moderate recall.

	Voice	Score
Precision	Active	99.4%
	Regular passive	73.7%
	Recipient passive	71.8%
Recall	Active	99.7%
	Regular passive	76.3%
	Recipient passive	57.8%

Table 5: Voice predictions for ditransitives with agent

In the no agent subset (i.e. the predominant active voice is excluded), the classifier achieves an accuracy of 81.31% outperforming the regular passive baseline about 20%. We then scored the precision and recall for the passive voices again (Table 6). Without the active voice, precision and recall for the recipient passive improve.

	Voice	Score
Precision	Regular passive	83.4%
	Recipient passive	77.7%
Recall	Regular passive	86.6%
	Recipient passive	73.1%

Table 6: Voice predictions for ditransitives without agent

6. Analysis

In the last section, we have shown that it is possible to predict the usage of the recipient passive by means of multiple contextual factors that mainly concern the properties of the predicate arguments. In this Section, we look at some of the features in more detail in order to see whether the model conforms to linguistic tendencies observed in previous work. Doing this, we have to be aware of the fact that our data has been automatically parsed and annotated, and therefore necessarily contains errors.

Table 8 shows the respective argument properties for all three voices. Looking at the pronominality of the arguments (feature "pronoun"), we observe that each argument (agent, recipient, patient) is most often pronominalized in

the voice where it is realised as the subject. Thus, the agent has a high chance of being pronominal in the active (39.6%), but is not likely to be realised as a pronoun in the passives voices (3.2% and 9.5%). This statistical tendency is in accordance with previous works observing a bias of the canonical subject position to be a slot for given actants (i.e. for those that can be expressed by pronouns) (Aissen, 1999; Bresnan et al., 2001). We find a similar tendency for the person feature. Thus, the agent is more likely to be a non-third person entity in the active than in the passive voices. Features that do not excatly match theoretical hypotheses are "Definiteness" (Def. in Table 8) and "Indefiniteness". We would expect that, for instance, the recipient is often indefinite in the active and regular passive, and less often in the recipient passive. However, the opposite is true (6% in recipient passive vs. 2.2% in active and regular passive). On the other hand, the definiteness features of the agent and patient seem to be consistent with theoretical predictions.

An important feature in our classification is the surface order of the arguments. This feature is not discussed in work on English alternations where the word order is fixed. By contrast, in German, the voice alternations interact with word order variation. In Table 8, the distribution of the "Precedes" feature shows that, for each voice, the subject precedes the objects most often (except of the recipient preceding the patient). Furthermore, the (optional) agent precedes the non-subject patient in recipient passive more often (43.5%) than the non-subject recipient in regular passive (26.3%). Actually, in German, it is strongly dispreferred to realise the agent directly after the finite auxiliary of the regular passive.,as in Example (16-a). However, this surface order is possible in the recipient passive, see Example (16-b).

(16) a. * Das Formular wird von der Karla dem Karl
          * The form   is   by the Karla the   Karl
          ausgefüllt.
          filled in.
          "The form is filled in by Karla for Karl."
      b. Der Karl bekommt von der Karla das Formular
          The Karl gets   by the Karla the form
          ausgefüllt.
          filled in.
          "The form is filled in by Karla for Karl."

For the distinction of the regular passive and the recipient passive, we had a feature encoding the presence of a prepositional agent. The distribution is given in Table 7 showing that the agent occurs twice as often in the recipient passive as in the regular passive. Thus, it seems that the agent is somehow more prominent in recipient passive than in regular passive. A possible explanation for this pattern could be found in the semi-grammaticalised status of the recipient passive. When the auxiliary *bekommen* still has a meaning reflex from its main verb, the agent is interpreted as the source of the possession transfer event.

7. Conclusion

We have presented a corpus-based study of the German recipient passive asking whether it is possible to predict the choice of the two meaning-equivalent passive voices in

	regular passive	recipient passive
agent	9.6%	21%

Table 7: Presence of a prepositional agent in passive voices

Feature	Role	Active	Pass.	Rec. Pass.
Pronoun	agent	39.6%	3.2%	9.5%
	recipient	40%	32.3%	45.7%
	patient	17.7%	24.9%	17.8%
¬3rd Pers.	agent	9.4%	0.4%	3.6%
	recipient	13.9%	7.9%	16.3%
3rd Pers.	agent	73.7%	95.1%	95.3%
	recipient	84.1%	89.9%	74.5%
Def.	agent	15.9%	49.3%	38.2%
	recipient	26.6%	31.1%	13.7%
	patient	18.4%	31.9%	22.2%
Indef.	agent	3.8%	3.5%	8.1%
	recipient	2.2%	2.2%	6%
	patient	16.7%	9.9%	23.3%
No Def.	agent	7.8%	18.3%	10.3%
	recipient	7.5%	9.6%	7%
	patient	22.6%	11.9%	12.9%
Proper N.	agent	7.9%	12.5%	13.6%
	recipient	7.4%	15.1%	10.5 %
	patient	6.4%	6%	10.3%
Common N.	agent	30.1%	74.2%	70.9%
	recipient	44.9%	43.9%	29.6 %
	patient	67.3%	60.1%	65%
Complex	agent	14%	22.7%	17.4%
	recipient	14.3 %	14.9%	10.8%
	patient	29.1%	23%	23.3%
Quantified	agent	1.1%	0.9%	0.4%
	recipient	1.7%	0.9%	2.6%
	patient	4%	2.9%	2.9%
Precedes	Ag > Rec	64.4%	26.3%	18.3%
	Ag > Th	65.6%	29.5%	43.5%
	Rec > Ag	22%	69.2%	75.7%
	Rec > Th	64.7%	39.7%	63.7%
	Th > Ag	15.5%	50.5%	45.9%
	Th > Rec	22%	37.6%	19.8%
Reentrant	Recipient	0.8%	0.6%	10.3%
	patient	1.9%	4.7%	1.7%

Table 8: Argument properties used in voice classification

German, regular passive and recipient passive. We created a specialised data set of occurrences of German ditransitive verbs. We parsed the sentences with the German LFG parser and extracted contextual features for each involved argument. Giving these feature sets as input to a maximum entropy classifier yield a good performance beating the active voice baseline.

This study illustrates that the German recipient passive alternation is subject to soft statistical constraints that can be modelled by means of multiple contextual factors.

8. References

Judith Aissen. 1999. Markedness and subject choice in optimality theory. *Natural Language and Linguistic The-*

ory, 17(4):673–711.

Judith Aissen. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory*, 21:435–483.

Jennifer Arnold, Thomas Wasow, Anthony Losongco, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76:28–55.

Joan Bresnan, Shipra Dingare, and Christopher D. Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG01 Conference*, pages 13–32, Stanford, CA. CSLI Publications.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2005. Predicting the dative alternation. In *Cognitive foundations of interpretation*, pages 69–94.

Joan Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell Publishers, Oxford, UK.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research*, pages 23–32, Budapest, Hungary.

Elke Diedrichsen. 2004. The german 'bekommen-passive' and rrg. In *Proceedings of the International Role and Reference Grammar Conference (RRG)*, Dublin, Ireland, October.

Gabriele M. Diewald. 1997. *Grammatikalisierung. Eine Einführung in Sein und Werden grammatischer Formen.* Niemeyer, Tübingen.

Peter Eisenberg. 1999. *Grundriss der deutschen Grammatik.* Metzlersche J.B. Verlagsb, Stuttgart, 2. edition.

Martin Forst. 2007. Filling statistics with linguistics – property design for the disambiguation of german lfg parses. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

Maxim Pimanyonok. 2004. *Das bekommen-Passiv im Deutschen.* GRIN Verlag.

Karin Pittner and Judith Berman. 2004. *Deutsche Syntax.* Narr Dr. Gunter.

Rajakrishnan Rajkumar and Michael White. 2011. Linguistically motivated complementizer choice in surface realization. In *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop*, pages 39–44, Edinburgh, Scotland, July. Association for Computational Linguistics.

Christian Rohrer and Martin Forst. 2006. Improving coverage and parsing quality of a large-scale LFG for German. In *Proceedings of LREC-2006*, Genoa, Italy.

Sandra Thompson. 1987. The Passive in English: A Discourse Perspective. In Robert Channon and Linda Shockey, editors, *In Honor of Ilse Lehiste*, page 497511. Foris, Dordrecht.

Sina Zarrieß, Aoife Cahill, and Jonas Kuhn. 2011. Underspecifying and predicting voice for surface realisation ranking. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1007–1017, Portland, Oregon, USA, June. Association for Computational Linguistics.

Sina Zarrieß, Aoife Cahill, and Jonas Kuhn. To appear. To what extent does sentence-internal realisation reflect discourse context? In *To appear in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012).*