Lang Resources & Evaluation (2012) 46:565–600
DOI 10.1007/s10579-011-9156-x

ORIGINAL PAPER

Evaluating automatic annotation: automatically detecting and enriching instances of the dative alternation

Daphne Theijssen · Lou Boves · Hans van Halteren · Nelleke Oostdijk

Published online: 6 July 2011

Abstract In this article, we automatically create two large and richly annotated data sets for studying the English dative alternation. With an intrinsic and an extrinsic evaluation, we address the question of whether such data sets that are obtained and enriched automatically are suitable for linguistic research, even if they contain errors. The extrinsic evaluation consists of building logistic regression models with these data sets. We conclude that the automatic approach for detecting instances of the dative alternation still needs human intervention, but that it is indeed possible to annotate the instances with features that are syntactic, semantic and discourse-related in nature. Only the automatic classification of the concreteness of nouns is problematic.

Keywords Automatic annotation · Intrinsic and extrinsic evaluation · Syntactic alternation · Dative alternation · Logistic regression

1 Introduction

Much effort has been—and continues to be—put into developing corpora to provide linguists with suitable data in sufficient quantities to perform their research. Still, for many types of research the availability of data remains an issue: even when

D. Theijssen (✉) · L. Boves · H. van Halteren · N. Oostdijk
Department of Linguistics, Radboud University Nijmegen, Erasmusplein 1,
6525 HT Nijmegen, The Netherlands
e-mail: d.theijssen@let.ru.nl

L. Boves
e-mail: l.boves@let.ru.nl

H. van Halteren
e-mail: b.vanhalteren@let.ru.nl

N. Oostdijk
e-mail: n.oostdijk@let.ru.nl

numerous corpora are available, most of these are too small and/or have not been annotated with the required information. This means that linguists often have to extend the data and, while doing so, somehow have to provide the necessary annotations. This often involves costly manual labour. It might also involve acquiring copyright for the new data, since, ideally, the additional data and annotations should be made available to other researchers: only then will it be possible to verify any results of experiments based on these data. A possible approach to creating sufficiently large sets of suitable data that can also be accessed by other researchers is to make use of already existing corpora and provide what additional linguistic information is required automatically, employing computational tools. In this article, we address the question: Is data that is obtained and annotated automatically suitable for linguistic research, even if the data may contain a certain proportion of errors? We investigate this by focussing on a specific linguistic task considering syntactic alternation: modelling the dative alternation.

In syntactic alternations, speakers and writers have several syntactic options that are equally grammatical, but where usually one particular variant is considered to be more appropriate in a given context. For instance, speakers of English can choose between the *s*-genitive, as in *John's dog*, and the *of*-genitive, in *the dog of John* (e.g. Rosenbach 2003). Researchers have suggested many factors that may influence the choice of a construction in syntactic alternations. These are syntactic (complexity, person, number, etc.), semantic (e.g. animacy, concreteness) and discourse-related (e.g. discourse givenness, structural parallelism) in nature. Research directed at syntactic alternations therefore needs various levels of annotation in order to be successful.

To address the question of whether automatic annotation of data affects linguistic models, we consider a well-studied type of syntactic alternation: the dative alternation. The dative alternation occurs in various languages, for example in English (e.g. Bresnan et al. 2007), Dutch (e.g. Colleman 2006), Greek (e.g. Anagnostopoulou 2005), Spanish (e.g. Beavers and Nishida 2010) and Brazilian Portuguese (e.g. Gomes 2003). In this article, we take the dative alternation with *to* in English as a case study.[1] The two syntactic constructions available in this alternation are the prepositional dative construction with *to* (example 1) and the double object construction (example 2).

1. The evil queen gives the poisonous apple to Snow White.
2. The evil queen gives Snow White the poisonous apple.

The English dative alternation has been studied by many researchers (e.g. Quirk et al. 1972; Collins 1995; Gries and Stefanowitsch 2004). More recently, a line of research has emerged which successfully combines the largely complementary theories advanced by these authors. Bresnan et al. (2007) and Theijssen (2010) have built logistic regression models that predict the construction used on the basis of features at different levels: the animacy of the recipient (*Snow White* in the example), the concreteness of the theme (*the poisonous apple*), the definiteness of the recipient and the theme, the discourse givenness

[1] English also allows alternations with prepositions other than *to*, for instance with *for* (also referred to as the 'benefactive alternation'). In this article, 'dative alternation' and 'prepositional dative' always refer to the variant with *to* only, unless explicitly indicated otherwise.

of the recipient and the theme, the length difference between the theme and the recipient (in terms of the number of words), the number of the recipient and the theme, the person of the recipient and the pronominality of the recipient and the theme. The regression models can predict over 90% of the instances correctly, using manually annotated data. The models show that animate objects tend to be placed before inanimate objects, concrete before abstract, definite before indefinite, discourse-given before discourse-new, shorter before longer, etc.

The previous research has resulted in two data sets that have been created in a way many linguists create their data sets: Researchers extracted as many candidates as possible from corpora that contain manually checked syntactic parses. All candidates were manually checked and manually annotated with the features required (Bresnan et al. 2007; Theijssen 2010). We employ these *traditionally established* data sets in different ways. We will use the data set in Theijssen (2010) as a development and analysis set: to optimise the algorithms, and to evaluate the errors made by them. We will refer to this set as Ice-Trad since the instances were extracted from the British component of the ICE corpus (ICE-GB, Greenbaum 1996). The data set established by Bresnan et al. (2007) is next used as a separate test set, for the purpose of quantitative evaluation only. It is taken from the Switchboard corpus of American telephone dialogues (Godfrey et al. 1992), and will be referred to as Swb-Trad from now on.

The goal of this paper is to evaluate the quality of two data sets that we extract from the same corpora, but automatically: Ice-Auto and Swb-Auto. The procedure for automatically creating annotated data sets consists of two steps: finding instances of the dative alternation, and enriching them with the desired information. Both steps will be elaborately described further on in this article, and they are evaluated independently in *intrinsic* evaluations. In order to establish the effect of the automatic procedure on our conclusions in linguistic research, we also need an *extrinsic* evaluation. We therefore evaluate the suitability of Ice-Auto and Swb-Auto by building new regression models on these sets and comparing the results to the models found for Ice-Trad and Swb-Trad.

The remainder of this article is organised as follows: Sect. 2 give a brief description of the two traditional data sets Ice-Trad and Swb-Trad. The automatic detection of instances is next described and intrinsically evaluated in Sect. 3, the automatic annotation of these instances in Sect. 4. The extrinsic evaluations are presented and discussed in Sect. 5. A general discussion and our final conclusion can be found in Sects. 6 and 7.

2 Traditional data

As development data, we employ the traditional data set in Theijssen (2010), Ice-Trad. It consists of instances found in the British component of the ICE Corpus (ICE-GB, Greenbaum 1996). The ICE-GB corpus contains spoken and written British English in various genres, as can be seen in Table 1. The corpus can be obtained from the Survey of English Usage.[2]

[2] See http://www.ucl.ac.uk/english-usage/projects/ice-gb.

Table 1 Number of double object constructions (d.obj), prepositional dative constructions (p.dat), total number of constructions (tot), and verb types (vb) per subgenre in the ICE-GB Corpus

Medium	Genre	Subgenre	d.obj	p.dat	tot	vb
W (200)	Non-printed (50)	Non-prof. writing (20)	11	3	14	7
		Correspondence (30)	93	32	125	29
	Printed (150)	Academic writing (40)	19	13	32	13
		Non-acad. writing (40)	35	13	48	15
		Reportage (20)	30	18	48	18
		Instructional writing (20)	25	10	35	8
		Persuasive writing (10)	8	8	16	6
		Creative writing (20)	45	9	54	18
			266	*106*	*372*	*51*
S (300)	Dialogues (180)	Private (100)	151	49	200	20
		Public (80)	116	41	157	25
	Monologues (100)	Scripted (70)	101	33	134	22
		Unscripted (30)	26	20	46	11
	Mixed (20)		12	9	21	9
			406	*152*	*558*	*43*
Total (approx. 1M words)			672	258	930	65

The number of corpus samples in the subgenres is given in brackets (each containing approx. 2,000 words)

The italics are used to indicate the subtotals per medium (spoken/written)

The procedure for establishing Ice-Trad was as follows. First, candidate sentences were automatically extracted from the corpus, making use of its manually checked syntactic parses. Next, all candidate sentences found were manually checked. This was especially necessary for the prepositional dative instances, since the syntactic annotation of the ICE-GB corpus does not distinguish between different types of prepositional phrases at the clause level. This means that sentences like example 3, in which the prepositional phrase is a locative, are also found, and should be filtered out manually. The resulting data set contains 930 instances in spoken and written British English. The number of instances and of different verb types in each subgenre of the corpus can be found in Table 1. The majority construction is the double object construction, with a relative frequency of 72.3% (672/930). With respect to medium, the proportion of instances in spoken data is highest: 60.0% (558/930).

3. Fold the short edges to the centre. (ICE-GB W2D-019 144:1)

As a test set, we employ the traditional Switchboard set (Swb-Trad), a set of 2,349 instances, being a corrected version of the original set described in Bresnan et al. (2007).[3] The Switchboard corpus consists of spoken telephone dialogues in American English (Godfrey et al. 1992) and can be obtained from the Linguistic

[3] We thank Prof. Joan Bresnan for sharing this data set with us.

Table 2 Number of double object constructions (d.obj) and prepositional dative constructions (p.dat) for the 10 most frequent verbs in Ice-Trad and Swb-Trad (*cost* and *take* are not included at all in Ice-Trad)

Verb	Ice-Trad				Swb-Trad			
	d.obj		p.dat		d.obj		p.dat	
	nr	perc	nr	perc	nr	perc	nr	perc
Give	377	**85.7**	63	14.3	1078	**85.8**	179	14.2
Offer	32	**76.2**	10	23.8	20	**66.7**	10	33.3
Send	51	**68.0**	24	32.0	89	**64.0**	50	36.0
Show	43	**81.1**	10	18.9	46	**86.8**	7	13.2
Teach	7	**100.0**	0	0.0	58	**95.1**	3	4.9
Tell	73	**98.6**	1	1.4	113	**96.6**	4	3.4
Bring	7	**70.0**	3	30.0	19	44.2	24	**55.8**
Cause	5	38.5	8	**61.5**	8	**80.0**	2	20.0
Cost	0	0.0	0	0.0	137	**100.0**	0	0.0
Do	10	50.0	10	50.0	25	**92.6**	2	7.4
Lend	9	**52.9**	8	47.1	2	**66.7**	1	33.3
Pay	8	32.0	17	**68.0**	83	**58.9**	58	41.1
Take	0	0.0	0	0.0	2	3.4	56	**96.6**
Sell	1	8.3	11	**91.7**	30	40.0	45	**60.0**

The percentages in boldface are those that are above 50%

Data Consortium.[4] For details about the extraction of this data set, we refer to Bresnan et al. (2007). Swb-Trad consists of 1,850 instances with a double object construction (78.8%), and 499 with a prepositional dative construction. The number of different verb types is 38.

Some verbs have a clear bias towards one of the two constructions, as can be seen in Table 2. In the top part, we see the verbs that show a bias towards the double object construction: *tell, teach, give, show, offer* and *send*. The bottom shows that the verb *sell* prefers the prepositional dative construction. For the verbs in the middle, the alternation differs in the two data sets (*lend, do, cause, pay* and *bring*) or the verb only occurs in one of these data sets (*cost, take*). The Table thus reveals that the two data sets were established with different conditions: Swb-Trad includes instances with *cost* and *take*, while they were not kept as instances in Ice-Trad.[5]

Both Ice-Trad and Swb-Trad have been manually annotated, for 12 features taken from Bresnan et al. (2007), as summarised in Table 3. Most features describe characteristics of the theme and the recipient in the construction. Some of the characteristics, however, are only established for either the recipient *or* the theme. Previous research (Bresnan et al. 2007) has shown that for some features, the theme and recipient are too biased towards one of the two values: themes are hardly ever

[4] See http://www.ldc.upenn.edu.

[5] The verb *cost* was left out because two linguists (the first and fourth author) judged that no alternation is possible. The verb *take* either occurred in prepositional dative constructions that were locative, or in double object constructions that were judged to alternate with the preposition *of*.

Table 3 Features and their values

Name	Feature	Values	Description
AnRec	Animacy of rec	a, in	Human or animal, or not
ConTh	Concreteness of th	c, a	With fixed form/space, or abstract
DefRec, DefTh	Definiteness of rec&th	d, in	Definite pronoun, proper name or noun preceded by a definite determiner, or not
GivRec, GivTh	Disc. givenness of rec&th	g, new	Mentioned or evoked $\leq$ 20 clauses before, or not (new)
LenDif	Length difference	−3.4 to 4.2	ln(#words th) − ln(#words rec)
NrRec, NrTh	Number of rec&th	sg, pl	Plural in number, or singular
PrsRec	Person of rec	l, non	Local (1st or 2nd) person, or not
PrnRec, PrnTh	Pronominality of rec&th	p, non	Headed by a pronoun, or not

th Theme, *rec* Recipient

animate and/or in first or second person (*me* or *you*), while recipients are hardly ever abstract. For this reason, the features Animacy of Theme, Person of Theme, and Concreteness of Recipient have been excluded in our previous research (Theijssen 2010) and are excluded in this article as well.

The manual annotations of Ice-Trad were done by the first author, following the annotation instructions provided in the "Appendix". The definitions are as close as possible to the descriptions used for Swb-Trad (Bresnan et al. 2007). In order to establish the quality of the data sets, we had an extra human annotator annotate subsets of the data sets. For Ice-Trad, the third author annotated 10 items that were randomly selected, after which he was provided with feedback about his annotations. After this short training session, he annotated 40 additional instances, on which κ scores were established. Only the inter-annotator agreement for Animacy of Recipient was below 0.75 (0.63). This unexpectedly low κ score was the result of only three disagreements, all concerning groups of people that could be interpreted either as institutions (being inanimate) or as groups of individuals (being animate). They have such a great impact on the κ score because there is a great bias towards animate recipients in the 40 items. For Swb-Trad, the first author annotated a subset of 30 items. The κ scores between these annotations and the original annotations by Bresnan et al. (2007) were 0.78 or higher for all features, which shows a high overall agreement. The individual κ scores per feature per data set will be provided later in this article, in Table 6 (being the results of the automatic feature extraction in Sect. 4)

3 Automatic detection of instances in a corpus

As mentioned in Sect. 1, the first step towards automatically obtaining data sets for studying the English dative alternation (Ice-Auto and Swb-Auto) is to detect instances automatically.

3.1 Related work

The dative alternation, together with other *diathesis* alternations,[6] has been the topic of interest for a number of researchers in the field of automatic lexicon learning, or more specifically: ‘verb classification’. Their goal has been to automatically induce possible verb frames[7] from corpora (for comprehensive overviews, see Schulte im Walde 2009; Korhonen 2009). Several approaches have been rather successful (e.g. Joanis et al. 2008; Schulte im Walde et al. 2008; Li and Brew 2008; Sun and Korhonen 2009), but many challenges are still to be met (Korhonen 2009). Only a few researchers have attempted to tackle the detection of actual instances of diathesis alternations automatically. Their work is shortly summarised below.

Lapata (1999) used the British National Corpus (BNC Consortium, 2007) to determine the frequency with which verbs occur in prepositional dative (with *to* and *for*), and double object constructions. First, she parsed the corpus with the shallow parser Gsearch (Keller et al. 1999) and extracted syntactic patterns that were potentially relevant. Next, she used a number of heuristic rules to divide the candidate patterns into relevant and irrelevant instances. The procedure was evaluated by comparing against manual annotations. For the double object construction (3,000 manually annotated candidates), the precision of the heuristics was approximately 89.8%, while for the prepositional dative construction with *to* (994 candidates), it was 77.3%. There is no information about recall.

McCarthy (2001) used syntactic and semantic cues to find various syntactic alternations, including the dative alternation. She parsed parts of the written part of the BNC with a probabilistic chart parser and an LR (left-to-right) parser based on string analysis. Looking at the most prototypical subcategorisation frames for each verb, she found six dative verbs that occur freely with different themes and recipients: *award, give, hand, lend, offer* and *owe*. She concluded that for the detection of instances of the dative alternation (with *to* and *for*), it is sufficient to use syntactic information only.

Lapata and Brew (2004) detected semantic preferences of verbs in the BNC and used them as priors in a Naive Bayes verb classifier. They used over 5,000 manual verb classifications to test against. Although they also evaluated the performance on the individual tokens, their task is essentially different from ours: they classify the verb class of a particular instance, while we want to detect instances of a certain verb class. The same is true for Girju et al. (2005). Using the annotations available in the PropBank, they used a machine learning technique to assign verb classes to instances (tokens).

Grimm and Bresnan (2009) automatically extracted instances of the dative alternation from a POS-tagged version of the Brown family of corpora (Hinrichs et al. 2007), consisting of the written American English corpora Brown (1960s) and

[6] Diathesis alternations are alternations in which verbs systematically allow a choice between two verb frames (double object, prepositional dative) to express the same semantic roles (recipient, theme).

[7] Verb frames indicate what type of arguments a given verb can take. The definition of types depends on one’s goal, and can be syntactically and/or semantically motivated.

Frown (1990s), and the written British English corpora LOB (1960s) and F-LOB (1990s). They parsed the corpora with the Stanford dependency parser and used a Python script to extract sentences with the desired syntactic pattern and a dative verb. The sentences with complex syntactic structures (e.g. passives) were filtered out. The procedure was evaluated on a small random subset of 100 sentences with the verb *give* in the Brown Corpus. For this small set, the accuracy for automatically distinguishing datives from non-datives was 45.0%, the recall 93.8% and the precision 46.4%. Given the low precision, they manually checked all 6,759 candidates, resulting in a final set of 3,114 instances that they used for further analysis.

3.2 Our method for automatic instance detection

For the automatic detection of instances, we use five steps that are performed in sequence:

1. Establishing a list of dative verbs
2. Extracting all sentences with these verbs from the corpus
3. Parsing the sentences with the FDG parser
4. Extracting candidates from the parses
5. Filtering the candidates with heuristic rules.

In the first step, we compile a list of dative verbs. This is not a necessary step, since we could simply include all dative constructions that the syntactic parser detects. Since we plan to use the automatic approach to case detection on very large corpora in the future, it is more efficient to first make a selection of potentially relevant sentences or utterances on the basis of a list of verbs.[8] The parsing, extracting and filtering then only needs to be applied to the retrieved sentences or utterances. Steps three to five are based on approaches in previous research. We use a syntactic parser to automatically extract potentially relevant instances like McCarthy (2001) and Grimm and Bresnan (2009). The candidates are filtered with the help of linguistic rules based on those in Lapata (1999).

In step one, we consider all verbs suggested in at least two of the following linguistic resources: the dative alternation verbs in Levin's verb classification (1993), the prepositional dative and double object frames in VerbNet (Kipper et al. 2000), the ditransitive verbs present in the ICE-GB corpus and the TOSCA lexicon (Oostdijk 1996), the verbs included in Bresnan et al. (2007), a list created by Johan Bos[9] and a list in an English Grammar Guide.[10] Many of the 264 verbs found are rather rare: 86 have a frequency below 1,000 in the 100-million-word British National Corpus (BNC Consortium, 2007). The occurrences of these verbs in the

[8] Of course this list should not be seen as static; language changes all the time, and new dative verbs emerge.

[9] Extracted from http://www.coli.uni-saarland.de/bos/atp/dtvs.html (which is no longer available): *ask, bring, buy, call, consider, demonstrate, describe, give, hand, leave, lend, offer, pass, promise, provide, send, serve, show, suggest, teach, tell.*

[10] See http://learning.cl3.ust.hk/english-grammar-guide/Verbs/Ditransitive_Verbs.htm.

Table 4 Final list of *dative verbs*, i.e. verbs that allow dative alternation and occur at least 1,000 times in the BNC

Accord	Cause	Flick*	Lower*	Promise	Serve	Take
Advance	Charge	Fling*	Make	Propose*	Ship*	Teach
Allocate	Concede*	Forbid	Offer	Quote*	Shoot*	Tell
Appoint*	Deal	Give	Owe	Read*	Show	Throw
Assign	Deliver	Grant	Pass	Recommend*	Sign*	Toss*
Award	Deny	Guarantee	Pay	Refuse	Signal*	Trade*
Bear	Dictate*	Hand	Permit	Repay*	Sing*	Vote
Bid	Do	Issue	Play	Return*	Slide*	Wish
Bounce*	Drop*	Kick*	Pose	Roll*	Slip*	Write
Bring	Extend*	Leave	Prescribe*	Sell	Submit*	Yield*
Carry*	Feed	Lend	Present	Send	Supply*	

Verbs marked with * are not recognised as allowing dative verb frames by the FDG parser

BNC are often in syntactic contexts that are not dative constructions. Since the eventual goal of the automatic detection of dative instances is to prevent data sparseness in future data sets, and since the verb itself is a feature in the statistical analyses, we want to exclude such low-frequency verbs.[11] This means we remove the aforementioned 86 verbs that occur fewer than 1,000 times in the BNC (e.g. *fax*).[12] Next, we manually filter out the 102 verbs that alternate with a preposition other than *to* (e.g. *cook for*) and/or that allow only one of the two constructions (e.g. *inform*). The procedure results in the list of 76 dative verbs in Table 4.

For the second step, we extract all sentences with a dative verb which occurs in the final list. If a corpus contains POS tags, most of the times they have either been checked manually (as is the case for the ICE-GB corpus)[13] or established automatically with the help of a tagger that is trained on similar material. We use the POS tags in the corpus for a first filtering: We only extract sentences if they have a dative verb that is tagged as a verb in the corpus. This filtering is left out in the evaluation on Switchboard, where we only use the plain text in the corpus.

In step three, the sentences are fed to the Functional Dependency Grammar (FDG) parser, version 3.9, developed at Connexor (Tapanainen and Järvinen 1997). The parser outputs functional dependencies that represent the structural information within the sentence. Our motivation for choosing this parser is fourfold. First, the level of detail is sufficient for our purposes, and both dative constructions are recognized. They are explicitly marked as datives, making the extraction of candidates straight-forward. For most parsers, this is not the case: they either only

[11] We tested the effect of verb frequency by including it as a fixed effect in regression models for Ice-Trad and Swb-Trad. In both models, the effect of verb frequency was far from significant. We therefore believe that removing the low-frequency verbs is warranted.

[12] The threshold of 1,000 is based on our observations of the list of BNC frequencies and our intuitions about the subcategorisation frames in which these verbs may occur.

[13] Actually, the leaf nodes of the syntactic parses contain information that is similar to the result of POS tagging, and these syntactic parses have been checked manually. We will refer to this information as 'POS tags' in the remainder of this article.

mark explicitly the double object construction (e.g. Stanford parser, Minipar, Link Grammar), or provide no function labels at all (e.g. Charniak). Second, the FDG parser can be used 'off-the-shelf', i.e. there is no need for training prior to applying it to data. This was an important motivation because of the small size of ICE-TRAD (only 930 instances, taken from a corpus of only 1 million words), which we use for developmental purposes. The Bikel parser (Bikel 2002), which does seem to distinguish the two dative constructions, could not be employed because it needs training. The FDG parser has been developed using approx. 100 million words in various kinds of texts—news articles, technical and legal documents, literature, discussion forum texts, transliterations, etc.—aiming for general use. Third, the parser does not need large computer capacity, and is quite fast in processing large amounts of data. Fourth, initial tests with the demo version of the parser[14] showed that it was able to deal with dative constructions with various verbs, and the parser was able to deal with complex sentences. A disadvantage of the parser is that 31 of the 76 dative verbs in Table 4 are not in the lexicon as being dative verbs (and cannot be added as such by users).

In the fourth step, we extract candidates from the syntactic parses. The parser generates one parse per sentence. In case a word is ambiguous, all possible functional and part-of-speech (POS) tags are provided, but it is always assigned only one relation. In dative sentences, the theme (*the poisonous apple* in the example) is labelled by the parser as an object ('obj') of the verb, while the recipient (*Snow White*), or the preceding *to* in the prepositional dative variant, is recognised as its dative complement ('dat'). We save all clauses in which one dative verb has both an object and a dative.

The fifth step consists of filtering the candidates found with heuristic rules. We distinguish between two types of filtering.

3.2.1 First filtering

Following Theijssen (2010), we exclude candidates that have at least one of the following features: (1) the theme or recipient is a clause, (2) the clause is in passive voice, (3) the verb is imperative, (4) the theme or recipient precedes the verb, (5) the verb is phrasal (e.g. *I'll send you out that*), (6) the clause is interrogative, (7) recipient and theme are reversed with respect to the expected order (e.g. *I give to him a letter*), (8) the theme is an adjective, (9) the theme or recipient is empty, (10) the clause is a fixed expression (e.g. *I'll tell you what*), (11) there is more than one verb, theme or recipient (e.g. *I gave it to her and to him*). Most of these filters are used to prevent the influence of other types of syntactic variation than those of interest in this research (passive versus active voice, declarative versus interrogative mode, the placement of adverbials, etc.). Some are used to make sure that the features we want to apply later are applicable (e.g. it is not possible to establish the concreteness of the theme if it is a clause, not a noun phrase). We use a Perl script to apply these filters automatically, making use of the automatic parses and a manually established list of fixed expressions. This list is based on the observations made

[14] See http://www.connexor.eu/technology/machinese/demo/syntax.

during the manual checking of the data set extracted from the ICE-GB corpus (Ice-Trad).

3.2.2 Second filtering

Obviously, syntactic parsing is not the easiest task, and parsers always make mistakes. This is certainly also the case for the two dative constructions, since they are often structurally ambiguous. For the double object candidates, the difficulty lies in word sequences that are difficult to split into phrases, like *the holy water* in example 4. For prepositional dative candidates, the problem is that the prepositional phrase can be either attached to the verb or the noun (e.g. *to parliament* in example 5). These problems are even worse in automatic parsing, since even candidates that are completely unambiguous for humans, are still ambiguous for the parser since it lacks world knowledge.

4. He gave the holy water.
5. They give access to parliament.

Given the fact that parsers make errors, we have a final step in which we remove candidates that have been falsely accepted due to errors in the parses. Following Lapata (1999), we formulate a number of heuristic rules to filter out these candidates. The rules we apply are based on Lapata's work and our observations of Ice-Trad. For some of the rules we need POS tags. For Ice-Auto, we use the POS tags available in the corpus; for Swb-Auto, we employ the POS tags provided in the automatic parse.

For both constructions, we remove all instances where the recipient or theme lacks the presence of a pronoun or noun. In these cases, the recipient or theme instead consists of a numeral, adjective or adverb, e.g. *a hollow* in example 6.

6. She gave [a hollow]$_{\text{Rec}}$ [laugh]$_{\text{Th}}$

For the double object constructions, there are more patterns that are likely to be the result of parse errors, or that represent structures that we do not consider instances of the dative alternation. More specifically, we filter out all candidates in which

- the last word of the recipient and the first word of the theme are proper nouns (e.g. *give John Smith*)
- the last word of the recipient is a possessive (e.g. *give Mary's money*)
- the last word of the theme is a reflexive pronoun (*give it yourself*)
- the verb is *make*, and both recipient and theme are persons in WordNet (Fellbaum 1998) (e.g. *make him king*)
- the verb is *take*, and the theme is a time noun in WordNet (e.g. *takes me an hour*)
- the recipient and theme together are likely to be one phrase (e.g. *write the professional letters*)

For the last rule, we need to establish whether the recipient and theme together are likely to be a single object. If the recipient ends in and the theme starts with at least one noun, we take the maximum sequence of ‘nouns’. This sequence not necessarily consists of real nouns only, since there may be errors in the POS tags. For instance, if we use the POS tags provided by the FDG parser (as we do for the Switchboard data), the parser could recognise a dative construction in *write the professional letters*. As a result, the word *professional* is tagged as a noun while it is in fact an adjective. We filter out such word sequences by first checking if it is present in a compound dictionary derived from WordNet (following Lapata 1999). If it is, the candidate is rejected (e.g. *holy water* in example 7).

If it is not, we use a corpus-based approach to establish the probability that the two or three words together form a single phrase (e.g. *sea water* and *priests water* in examples 8 and 9 respectively). For this, we slightly adapt the approach in Lapata (1999), using Daudaravičius and Marcinkevičiene’s *gravity* measure (2004), as suggested in Gries (2010), instead of the log-likelihood ratio. Gravity (G) not only takes into account the token frequencies of the separate words A and B and that of the sequence A-B, but also the number of possible word types before B and after A:[15]

$$G = \log\left(\frac{F_{AB} * N_b}{F_A}\right) + \log\left(\frac{F_{AB} * N_a}{F_B}\right), \tag{1}$$

in which F_{AB} is the token frequency of the combination A–B, F_A the frequency of word token A, F_B the frequency of word token B, N_a the number of possible word types before B, and N_b the number of possible word types after A. The values are based on the British National Corpus (BNC Consortium, 2007). Using this formula, we calculate the gravity between two nouns in the sequence. If the gravity is above the suggested threshold of 5.5 (Daudaravičius and Marcinkevičiene 2004), the noun sequence is probably a single phrase and we thus reject the candidate (e.g. *give the sea water* in example 8). Else, we keep the candidate (e.g. *give the priests water* in example 9). For three-word sequences A–B–C (e.g. *priests holy water* in example 10), we first decide how to split it into two parts by establishing the gravity between the two possible pairs (A–B and B–C). The pair with the highest gravity is next used as input to the formula, together with the remaining word (A–B and C, or A and B–C).

7. He gave the holy water.
8. He gave the sea water.
9. He gave the priests water.
10. He gave the priests holy water.

For the prepositional dative construction, we exclude all instances where the recipient is a location in WordNet (e.g. *bring him to school*). Also, we remove the instances where the prepositional phrase is likely to be the complement of the theme rather than the verb (e.g. *give access to the garden*). We again employ the aforementioned gravity measure to establish this. Since the BNC contains no syntactic annotations at the level required, we first parse the BNC with the FDG

[15] The words ‘types’ and ‘tokens’ refer to the counts of unique words and of all words, respectively.

Table 5 Results of automatic case detection for both the development/analysis data (ICE) and test data (SWB)

	ICE	SWB
Number of candidates found by parser	1,674	5,087
Number of candidates after 1st filtering	1,111	3,356
Number of candidates after 2nd filtering (AUTO)	889	2,694
Number of candidates in both AUTO and TRAD	619	1,292
Precision (%)	69.6	48.0
Recall (%)	66.6	55.0
F-score (%)	68.1	51.2

parser. We then use the PP-attachment in the parses to calculate the gravity between the verb and the recipient, and between the theme and the recipient.

3.3 Results

We applied the approach described in the previous section to ICE-GB and Switchboard. The number of candidates found in both corpora are presented in Table 5. The table also shows the precision, recall and F-score of the sets after the second filtering (ICE-AUTO and SWB-AUTO), when comparing them to ICE-TRAD and SWB-TRAD.

In ICE-AUTO, 62.9% (559/889) of the instances are from the spoken part of the corpus, which is not significantly different from the 60.0% in ICE-TRAD ($\chi^2 = 1.47$, df = 1, $p > 0.20$). The majority construction is the double object construction, comprising 73.1% (650/889) of the instances (which again is not significantly different from ICE-TRAD: 72.3%, $\chi^2 = 0.13$, df = 1, $p > 0.70$). In SWB-AUTO, the proportion of double object constructions is significantly different from SWB-TRAD: 81.9% (2,206/2,694), compared to 78.8% ($\chi^2 = 7.61$, df = 1, $p < 0.01$).

3.4 Discussion

When we look at the precision, recall and F-score for both data sets, we see that the scores for ICE-AUTO are much higher than those for SWB-AUTO. In general, parsers have more difficulties with spoken material than written material, because it often contains disfluencies, corrections and unfinished clauses. We also find a trend for this within the ICE-GB data: The precision for the spoken instances in the ICE data is 67.4% (377/559), while it is 73.3% (242/330) for the instances in written English ($\chi^2 = 3.13$, df = 1, $p < 0.10$).

The approach is quite successful on the development data: with the help of the filtering rules, our approach outperforms the precision reached on the Brown corpora by Grimm and Bresnan (2009): 69.6% compared to 46.1%. The recall of our approach, however, is much lower: 66.6% compared to 93.8%. Combining precision and recall, we reach an F-score of 68.1% on the ICE-GB data, which is higher than the 61.8% obtained by Grimm and Bresnan (2009). When we compare our performance on the Switchboard data to Grimm and Bresnan (2009), we see that the precision we reach (48.0%) is comparable to the precision they reach on the Brown corpora (being 46.1%). Their F-score is much higher (61.8% compared to our 51.2%), however, because of their better recall. It is clear that the spoken data in the test set (Swb-Auto) is problematic for our approach.

Let us now consider the errors made on the development set (Ice-Auto). In order to gain insight into the possible improvements of the approach, we manually classified the 270 candidates in Ice-Auto that are not present in Ice-Trad:

- 131 (48.5%) of them are found because the FDG parser incorrectly recognised a dative construction. These sentences are not part of Ice-Trad because the syntactic annotations of these sentences in the ICE-GB corpus do not contain a construction that could be dative. These errors are thus in fact parse errors, which we are unable to solve (at least in the scope of this article).
- 125 (46.3%) are found both in the ICE-GB annotations and the automatic parses. In Ice-Trad, these were filtered out automatically (using the syntactic annotations in the corpus) or manually. The procedure is different for Ice-Auto: the automatic filtering is now based on the automatically obtained FDG parses, and the manual filtering is replaced by the filtering rules. For 125 instances, the FDG parser thus should have indicated that these constructions are irrelevant, or the filtering rules should have filtered them out. In 33 sentences, the prepositional phrase indicates a location, amount, time or degree, in 22 the verb is phrasal, 17 are fixed expressions, 15 are imperatives or interrogatives, 12 have an object that is split up or incomplete, 11 have clausal objects, and 15 are irrelevant because of other reasons. Where we manually checked these properties for Ice-Trad, we have performed no such checking for Ice-Auto.
- 14 (5.2%) actually contain relevant dative constructions. Of course, manually checked annotations are also error-prone (e.g. Nancarrow and Atwell 2007). The fact that these 14 instances are not part of Ice-Trad exemplifies this, since they are missed due to errors in the annotations in the ICE-GB corpus. Most of these instances (11) were prepositional dative constructions in which the prepositional phrase was incorrectly attached to the theme, not to the verb.

The division shows that of the 889 candidates found automatically, 256 are not relevant instances of the dative alternation. The FDG parser thus reached a precision of 71.2% on recognising the two objects in dative constructions. This precision is much lower than the general precision the parser reaches on linking subjects and objects: 93.5% on texts from the Maastricht treaty and 96.5% on foreign news texts.[16] Of the remaining 633 candidates, not all are present in Ice-Trad, but they are all instances of the dative alternation. The effect of the 256 irrelevant cases will become evident in Sect. 5.

[16] These figures were established by Connexor Oy in December 2005, after which only minor changes have been applied to the parser.

The 311 instances not found automatically can be subdivided as follows:

- 206 (66.2%) are due to errors in the automatic parses. Of these, 10 have a verb that is listed in our list of dative verbs (Table 4), but which is not marked as such in the parser lexicon (e.g. *read*, indicated with the asterisk in Table 4). So, the fact that 31 of the 78 dative verbs in our list are not stored as such in the parser lexicon, eventually causes only 10 instances to be missed.
- 43 (13.8%) are falsely filtered out in the second filtering. In 20 the gravity measure unjustly indicated that the recipient should be interpreted as a postmodifier of the theme. In 12 the recipient is interpreted as a location on the basis of WordNet, while it is not a location in the given context. In 11 the theme or recipient does not contain (or is taken not to contain) a (pro)noun, but only a numeral, adjective or adverb. Seeing the small number of errors per type of filtering, there is only little to gain by improving the filtering.
- 38 (12.2%) are not found because the verb is not in the list of dative verbs (Table 4). We have not employed this list to establish ICE-TRAD; we extracted all dative constructions regardless of the verb present, and manually checked whether the candidate was relevant.
- 24 (7.7%) are falsely filtered out in the first filtering: 10 are taken to have a clausal object, 6 are interpreted as expressions, 4 are taken to have split objects, 3 are considered passives or imperatives and 1 is taken to contain a phrasal verb. As with the second filtering, the numbers are too small to make investing in improvement worthwhile.

When we compare the recall we obtained with the FDG parser (66.6%) to the recall that the FDG parser reaches on linking subjects and objects in general, we see that the general recall of the FDG parser is much higher: 90.3% on texts from the Maastricht treaty and 95.4% on foreign news texts.[17]

Counting the number of words in the sentences in ICE-TRAD, we see that the instances that were found automatically have an average length of 21.5 words, which is significantly shorter than the average length of 25.7 words for the cases we could not find automatically ($t = -4.23$, $df = 548$, $p < 0.001$). Apparently, the parser has most difficulty identifying dative constructions in sentences that are relatively long and thus, presumably, more complex. Also, the division of the two construction types (double object and prepositional dative) differs significantly between the instances we have found in the ICE corpus and those we have missed. Of the cases we found automatically, only 19.4% is a prepositional dative, while this is true for 44.4% of the cases we have missed ($\chi^2 = 63.2$, $df = 1$, $p < 0.001$).[18] The attachment of prepositional phrases (PP-attachment) is a common problem in automatic parsing (e.g. Agirre et al. 2008). It is therefore not surprising that the parser has more difficulties with the prepositional dative variant than with the double object construction (cf. also Lapata 1999).

[17] Again, these figures were established by Connexor Oy in December 2005.

[18] As we will mention in Sect. 5, the difference in the distribution of the two constructions between ICE-AUTO and ICE-TRAD will not influence the extrinsic evaluation, because logistic regression is very robust against class imbalance.

4 Automatic annotation of instances found

Now we have described the automatic extraction of the instances, we can move on to the annotation of the instances. We include the features introduced in Sect. 2 Previous research has already shown that they play a role in the dative alternation (Bresnan et al. 2007; Theijssen 2010).

4.1 Method

The automatic extraction of the values for the twelve features is described below. Our aim is to obtain feature values that agree with the ones selected by the human annotator in Ice-Trad. We make use of the syntactic parses produced by the FDG parser. For some features, we consult WordNet (Fellbaum 1998). Most corpora contain POS information, so we also make use of the POS tags in the ICE-GB corpus. For the Switchboard data, we use the POS tags provided by the parser.

4.1.1 Animacy of recipient (AnRec)

Most researchers who have been successful in animacy classification of *English* nouns employ WordNet (e.g. Orăsan and Evans 2007; Baker and Brew 2008). We therefore employ WordNet as well, together with other resources. More precisely, we use three lists of animate words: (1) the nouns marked as person or animal in WordNet, (2) a list of company names found on the Internet[19] and (3) a short list of additional words, e.g. personal pronouns like *I* and *him*. Company names are thus deemed animate. Our assumption is that company names functioning as a recipient in a dative construction will mostly refer to the people working at this company (e.g. *BUPA* in example 11).

11. I mean two three weeks ago John Major made a speech to BUPA in which he said he wanted the private sector to be boosted. (ICE-GB S1B-039 64:1:C)

In this article, we simplify the problem of animacy classification of the recipient in two ways. First of all, we limit ourselves to the lemma of the syntactic head of the recipient, as found in the syntactic parse. Second, we classify the different types (not the different tokens), irrespective of context. This means that we always assign the same value for animacy to recipients that have as their syntactic head the same lemma. When this lemma is present in at least one of the three lists mentioned above (ignoring upper/lower case), we classify it as *animate*. All other recipients are deemed *inanimate*.

4.1.2 Concreteness of theme (ConTh)

In the dative alternation, the theme can either be "prototypically concrete" (Garretson 2003), i.e. having a fixed size or form in space (*She gave him **a book***), or

[19] The company names have been extracted from http://www.buyblue.org/alphalist.php (which is no longer available) and http://www.businessweek.com/1999/99_28/alphalist.htm. All names ending in Corp, Corporation, Co, Incorporated, Inc, Holding, Group were duplicated without this ending.

abstract (*She gave him* ***her love***). Concepts like *love* and objects like *book* are fairly straight-forward, but there are many difficult cases. Sometimes words have several senses. For instance, if a furniture salesman *shows you* ***a table***, he most likely refers to the concrete object standing in his showroom, while a researcher giving a presentation refers to the representation of information he or she has put on the slide. Furthermore, words in the same (or at least in a similar) sense can be used figuratively: when a waitress in a restaurant *shows you* ***a table***, you could say she is literally showing you a concrete object (a specific *table*), but what she means is not just a table, but a place to have dinner. In this situation, the *table* is arguably not prototypically concrete anymore.

The assumption that meaning depends on context is not new. The distributional hypothesis in Harris (1954) has led to a long line of context-based approaches in lexicon learning, many of which are semi-supervised or unsupervised. There are two main lines of research: (1) clustering semantically similar words (e.g. Rooth et al. 1999), and (2) extending existing lexicons through bootstrapping. In both, contextual features are used to find similarities. The clustering approach is not directly useful for us, since we want a binary, pre-defined, classification. A common method in bootstrapping is to start with a very simple lexicon, comprising a set of occurrences (tokens) of word types that are prototypical examples of the semantic class of interest (in our case concrete and/or abstract words). For this *seed set*, it is assumed that the word types have this class in (almost) all contexts, so you can use all tokens of this word type. In an iterative process, the seed set is extended with new word tokens that share properties with the tokens in the seed set. Many researchers use syntactic information for this purpose, for instance for classifying nouns into the lexical categories *building, event, human, location, time* and *weapon* (Riloff and Jones 1999; Thelen and Riloff 2002) or for detecting film titles (Kuijjer 2007).

In a previous study (Theijssen et al. 2011), we evaluated five different (semi-) automatic approaches for establishing the concreteness of nouns for the purpose of investigating the dative alternation. One approach used the MRC Psycholinguistic Database (Coltheart 1981) to find the concreteness value of a noun *type* (interval scale). Two approaches found the concreteness of a noun *sense* with the help of the WordNet hierarchy: by counting the number of nodes from the sense to the root (ordinal scale, following Changizi 2008), and by checking whether the sense was part of the *physical entity* subtree (binary value, following Xing et al. 2010). The final two approaches (both resulting in an interval scale score per noun *token*) were two variants of the bootstrapping approach in Thelen and Riloff (2002): one using all syntactic contexts, and one only dative contexts. The data used for bootstrapping was taken from the British National Corpus. The conclusion was that the first three approaches were hampered by the insufficient coverage of the lexical resources used (WordNet, MRC). The bootstrapping approaches were not very successful either: abstract nouns denoting ‘time’ (e.g. *minute, February*) and ‘quantity’ (e.g. *ton, inch*) received high concreteness scores. This shows that the selection of the seed set is not trivial, and that seeds are often not as suitable as one would expect.

Seeing the problems with the use of existing resources and with applying a bootstrapping approach, we decided to make use of our development data: ICE-TRAD.

We take the 619 instances in Ice-Trad that were also detected automatically, and establish a number of syntactic features for them. For each instance, we find the automatically obtained FDG parses and for the head of the theme, we extract the relations to its daughter nodes and to its mother node. Also, the POS tags and lemmas are retrieved, as well as the representation (upper/lower case, etc.). The found information is transformed into machine learning features. For instance, the features for *apple* (head of *the poisonous apple*) in the example sentence are:

- lemma of the focus word: *apple*
- upper (*U*) / lower case (*L*) and presence of non-alphanumeric characters (*S*) in the focus word: *L*
- POS tags of the focus word: *N_NOM_SG*
- relation with the mother node: *obj*
- relation with the mother node + its lemma: *obj;give*
- relation with the mother node + its POS tag(s): *obj;V_PRES_SG3*
- relation with the daughter node(s): *det, attr*
- relation with the daughter node(s) + its/their lemma(s): *det;the, attr;poisonous*
- relation with the daughter node(s) + its/their POS tag(s): *det;DET, attr;A_ABS*

After establishing the features for the themes in the instances, we applied various machine learning algorithms to classify the 619 instances in a tenfold cross-validation setting. We employed Weka (Hall et al. 2009) for a number of classifiers,[20] and libSVM (Chang and Lin 2001) for Support Vector Machines (SVMs). For all algorithms, only the features that occur at least three times in the training data were actually employed. The best results were obtained by the SVMs.

SVMs need tuning of three hyperparameters: (1) the kernel, which we limit to linear and RBF, (2) the cost *c*, for which we go through a grid of 2^{-12} to 2^{10} with steps of *2, and (3) the gamma *g* (only for the RBF kernel), for which we go through a grid of 2^{-10} to 2^{6} with steps of *2. The optimal hyperparameters are found in a tenfold cross-validation setting, and these are then used to build an SVM on all training data. When applying the classification to the 619 instances present in Ice-Trad, we perform leave-one-out: the tuning is done on 618 instances in tenfold cross-validation, the optimal settings are used to build an SVM on all 618 cases, which is then used to predict the 619th instance. When predicting data that does not overlap with the 619 instances in Ice-Trad, we use all 619 instances for tuning and training.

4.1.3 Definiteness of recipient and theme (DefRec, DefTh)

For establishing the definiteness of recipient and theme, we use the POS tags in the corpus or the parse. In order to establish what is the head of the object, and with which words it occurs (i.e. which words are its daughter nodes) we always use the dependency relations in the FDG parse.

When the head occurs with a definite article, we classify it as *definite*. The same applies to a head that is, or occurs with, a demonstrative, interrogative, relative or

[20] More specifically, we used Naive Bayes, Logistic, Multilayer Perceptron, Voted Perceptron, RBFNetwork, Ibk, AdaBoost, Bagging, SimpleMI, Jrip, DecisionTable, J48 and RandomForest.

possessive pronoun. Similarly, we consider *definite* heads that are a reciprocal, reflexive or personal pronoun, or a proper noun.

4.1.4 Discourse givenness of recipient and theme (GivRec, GivTh)

Automatically identifying discourse-new objects has received considerable attention of researchers working in the field of anaphora resolution. This is because the first step in anaphora resolution is recognizing which elements should be resolved, i.e. which elements actually refer to an item that has previously been mentioned (and thus is discourse-given).

Vieira and Poesio (Vieira 1998; Vieira and Poesio 2000; Poesio et al. 2004) used heuristics to establish which definite nouns are discourse-new. For example, one heuristic rule says that noun phrase heads that start with a capital (e.g. *The Iraq war*), or that refer to time (*the morning*) are discourse-new. Testing on 195 definite phrases, they reached a precision and recall of 77%.

The work by Vieira and Poesio has been extended in several ways. Some have added new heuristics (e.g. Bean and Riloff 1999). Others have extended the work to other types of nouns, not only the definite ones (e.g. Ng and Cardie 2002; Uryupina 2003). The use of different machine learning techniques for the task of detecting discourse-new objects has also received some attention (e.g Ng and Cardie 2002; Kabadjov 2007). Most researchers have employed the corpora created for one of the Message Understanding Conferences (MUC) for training and testing. The precision and recall are generally above 80%.

In the context of anaphora resolution, establishing the discourse givenness of an object affects the decision made further on in the discourse, since *new* objects will be *given* later in the discourse. Our task is considerably easier, since we only have to establish whether the recipient or theme is discourse-given or discourse-new. We therefore developed our own, much simpler, algorithm.

The approach we take is as follows. Given the fact that indefinite objects are mostly new to the discourse, we classify all indefinite objects as *discourse new*. For definite objects, we extract the head and its attributes from the FDG parse, and take the POS tags from the corpus or the parse. Definite objects of which the head is a personal pronoun, and of which the head is preceded by a demonstrative pronoun, are labelled *discourse given*. For the remaining definite objects, we check the preceding contexts, with a maximum length of 20 clauses (i.e. until the 20th preceding word that is tagged as main verb). If the head itself, or a synonym of the head is found within this preceding context, the object is considered *discourse given*. We use the synsets in WordNet to extract the synonyms. The remaining definite objects are given the value *discourse new*.

4.1.5 Number of recipient and theme (NrRec, NrTh)

Again, we employ the POS tags in the corpus or the parse. We use the FDG parse to identify the head of the object, and take the number provided in the POS tag. For heads of objects that have no information about number (e.g. *you*, which can be both), we assign the default value *singular*.

4.1.6 Person of recipient (PrsRec)

For this feature, we simply check whether the head of the recipient is *I, me, my, mine, myself, you, your, yours, yourself, yourselves, we, us, our, ours* or *ourselves*. If this is the case, the recipient is *local*, otherwise it is *non-local*.

4.1.7 Pronominality of recipient and theme (PrnRec, PrnTh)

For this feature, we again employ the POS tags in the corpus or the parse. We extract the head of the object from the FDG parse. If the head has a POS tag for (any type of) pronoun, the object is classified as *pronominal*. If not, it is *non-pronominal*.

4.1.8 Length difference (LenDif)

For length difference, we use a Perl function that we also used for Ice-Trad (see the "Appendix"). It counts the number of words in the recipient and the theme by splitting on white space, and takes the natural log of these lengths to smoothen outliers. The recipient length is then subtracted from the theme length (thus giving the log of the ratio between the lengths). The difference with Ice-Trad is that the input strings are now not the theme and recipient as found in Ice-Trad, but as found with the automatic approach (i.e. in the FDG parses).

4.2 Results

We intrinsically evaluate the automatic feature extraction by comparing the features values found to a gold standard. This gold standard consists of the manual annotations in Ice-Trad and Swb-Trad. For this reason, only the instances that are present in both the traditional and the automatic set can and will be included in this evaluation (619 for the ICE data, 1,292 for the Switchboard data).

For the only feature with an interval scale, length difference, we calculate the correlation coefficient between the values in the traditional set, and those in the automatic set. For the 619 instances in Ice-Trad and Ice-Auto, the correlation is 0.825. For 442 (71.4%) of the 619 instances, the feature value is exactly the same in both data sets. In 26 (4.2%) instances, the theme and recipient are equally long in one data set, but differ in length in the other. In only 8 (1.3%) instances, the polarity differs. For the remaining 143 (23.1%) instances, the same object is found to be longer than the other, but the length differences found differ.

For the 1,292 instances in Swb-Trad and Swb-Auto, the correlation is 0.635. The length difference has exactly the same value in 851 (65.9%) instances, is zero for only one of the two sets in 97 (7.5%) instances, differs in polarity in 11 (0.9%) instances, and only differs in the size of the length difference in 333 (25.8%) instances.

Table 6 The accuracy (Acc) of automatic feature extraction and the proportion of the majority class (Maj) in the traditional sets for the binary features

Feature	ICE				SWB			
	Acc	Maj	κ	Human κ	Acc	Maj	κ	Human κ
PrsRec	1.00	0.52	1.00	1.00	0.82	0.64	0.65	0.91
PrnRec	1.00	0.77	0.99	0.95	1.00	0.87	0.98	1.00
DefTh	0.99	0.65	0.97	1.00	0.97	0.70	0.94	0.93
DefRec	0.99	0.93	0.93	0.78	0.98	0.95	0.80	1.00
NrTh	0.98	0.86	0.92	0.88	0.98	0.80	0.93	1.00
PrnTh	0.98	0.88	0.89	0.84	0.97	0.84	0.91	1.00
NrRec	0.94	0.71	0.84	0.77	0.95	0.71	0.88	1.00
GivRec	0.91	0.82	0.69	0.95	0.95	0.86	0.79	0.80
AnRec	0.92	0.90	0.68	0.63	0.91	0.87	0.55	1.00
GivTh	0.87	0.83	0.59	0.80	0.90	0.82	0.68	0.78
ConTh	0.87	0.79	0.55	0.75	0.79	0.72	0.37	0.86

The fourth column indicates the κ score between the traditional and the automatic annotation of the instances in both sets. Human κ scores are provided in the fourth column

For the binary features, we calculated the classification accuracy and established the proportion of the majority value. The results are shown in Table 6. They show us two important things. First of all, most features are biased towards one of the two values. Over 90% of the recipients are definite, for instance. An exception is the person of the recipient where the division between *local* and *non-local* is much more balanced. Second, we see that the accuracies reached are all $\geq$79%.

Seeing that all accuracies and most majorities are above 79%, we have also established the κ statistic for inter-annotator agreement that discounts the prior probability that two annotators will agree. The two annotators in this case are the human annotator of the traditional set, and the extraction algorithm for the automatic set. Table 6 shows these κ scores, as well as those between two human annotators (as described in Sect. 2) Most of the κ values between the automatic and the manual annotations are quite similar to the κ scores between two human annotators. This is not the case for the features that are intuitively the most difficult (givenness, animacy and concreteness); they result in lower κ scores. For the other features, the κ scores are all above 0.65.

4.3 Discussion

When looking at the results for the ICE data (the development/analysis data), we see that for the animacy of the recipient, the κ score between the automatic extraction and the human annotations is very similar to that between two humans. Apparently, the simplifications in the automatic extraction have not influenced the quality of the extraction, and the resources we employed are quite reliable. One of those resources is WordNet. Of course, some noun head lemmas may have several senses and therefore occur not only as animal or person, but also in a different noun class.

This is the case for 12 of the 47 incorrectly classified instances: *authority* (twice), *dealer, face, man, master, mother, opposition, party, plant, subject* and *world*. With respect to simplifications, remember we had two of them: (1) limiting ourselves to the lemma of the syntactic head of the recipient, and (2) classifying the different types (not the different tokens), irrespective of context. An analysis of the 47 misclassifications reveals that only five are caused by the first simplification: four are animate but were labeled inanimate (*those who…, any of…, the rest of…, Mr …*) and one the other way around (due to a parse error in *the former Deputy Prime Minister's words*). The second simplification leads to sixteen errors in Ice-Auto. Three are nouns that are inanimate in the given context, but were labelled animate automatically: *world, nation* and *face*. The rest are incorrectly labelled inanimate, e.g. *few, it, jury* and *panel*.

The lowest κ scores between the automatic and manual annotations are for the concreteness of the theme, in both Ice-Auto and Swb-Auto. When we look at the 83 cases that are different in the ICE data, we see that fourteen are pronouns (*it, some* and *that*). It is not surprising that pronouns are difficult for the automatic approach; it depends even heavier on the context, since it needs to resolve to which antecedent the pronoun refers. In addition, there are twenty cases where the theme is something made of paper, e.g. *picture, piece of paper, card, voucher*. These are concrete in the manual annotation, but labelled abstract in the automatic approach. Apparently, these types of themes often share contextual properties with themes that are abstract.

Surprising is the rather low κ for the person of the recipient for the Switchboard data: 0.65. A quick look at the discrepancies shows that the annotations in Swb-Trad are more semantic in nature (whether the recipient is really physically part of the discourse), while we used a more syntactic definition (whether the recipient is in first or second person) in Ice-Trad and the automatic approach. Almost all of the differences were caused by the generic use of *you* (e.g. *it gives you energy*), which was labelled 'non-local' in Swb-Trad, and 'local' by us.

5 Extrinsic evaluation: Using the data in logistic regression models

The intrinsic evaluations in the previous sections have shown that the automatic detection of instances of the dative alternation may need improvement, but that the annotation of these instances seems promising. In order to establish the effect of the automatic procedure on our linguistic research, we need an extrinsic evaluation. The use of extrinsic evaluations is quite common in the field of Natural Language Processing (NLP), where the quality of the automatic annotations are tested in NLP applications like machine translation (e.g. Bod 2007) and question answering (e.g. Theijssen et al. 2007). Researchers in NLP now even question the use of *intrinsic* evaluations (e.g. Poibeau and Messiant 2008). For our extrinsic evaluation, we build a logistic regression model on the automatic data (Ice-Auto and Swb-Auto), and compare the effects in the model to a model built on the traditional data (Ice-Trad and Swb-Trad).

5.1 Method

Previous research has indicated that over 90% of the dative alternation can be correctly predicted with a logistic regression model that combines the features introduced previously (Bresnan et al. 2007; Theijssen 2010). More information about multivariate techniques such as regression can for instance be found in Izenman (2008).

As discussed in Theijssen (2010), there are at least six ways to build a logistic regression model for the dative alternation. One can choose between a mixed model, i.e. a model with a random effect, and a model without such an effect. Seeing the verb biases we presented in Table 2, we want to include verb as a random effect.[21] The second choice we have to make is the manner of feature (or *variable*) selection. Researchers have employed at least three different approaches to feature selection: (1) first building a model on all available explanatory features and then removing those that do not show a significant contribution (e.g. Bresnan et al. 2007), (2) sequentially adding the most explanatory feature (forward), until no significant gain is obtained anymore (e.g. Grondelaers and Speelman 2007), and (3) starting with a model containing all available features, and (backward) sequentially removing those that yield the lowest contribution (e.g. Blackwell 2005). Comparing all three options for the two data sets is beyond the scope of the present article. Seeing that our research is the closest to that in Bresnan et al. (2007), we follow their approach. We will thus build only one type of model: a mixed model with verb as a random effect, building it on all features, then removing all features that are not significant, and building a new model with only those features.

The ICE data sets differ from the Switchboard data sets in the sense that they contain both spoken and written material. Following Theijssen (2010), we include medium (spoken or written) as an additional feature, and add all interactions of medium with the twelve features of the previous section. This leads to a total number of 25 features. When removing non-significant main effects, we never remove those that are part of a significant interaction. For the Switchboard data, containing only spoken material, we include only the twelve main features.

5.2 Results for the ICE data

We build two regression models for the ICE data: (1) a model built on the 930 instances in Ice-Trad, and (2) a model built on the 889 instances in Ice-Auto. The model quality of these models (and an additional one that will be introduced later in this section) can be found in Table 7. The models fit the data well: the prediction accuracy is over 87%. This is significantly better than the majority baseline of always selecting the double object construction. Also, the concordance C is above 94% for all three models. In a tenfold cross-validation setting, the regression models show only a slight decrease in prediction accuracy and concordance C, which means there is hardly any overfitting.

[21] We use the function `lmer()` in the `lme4` package in `R` (R Development Core Team 2008).

Table 7 Prediction accuracy and concordance C for the model fit (*Acc, C*) and in tenfold cross-validation (*av. Acc, av. C*), for ICE-TRAD, ICE-AUTO and ICE-SEMI

Data set	Majority	N	Acc	av. Acc	SD	C	av. C	SD
ICE-TRAD	0.723	930	0.915	0.896	0.036	0.973	0.962	0.016
ICE-AUTO	0.731	889	0.880	0.871	0.045	0.947	0.933	0.025
ICE-SEMI	0.791	633	0.930	0.918	0.055	0.969	0.954	0.035

The majority baseline and the number of instances are also provided

Table 8 Significant features in the model built on the 930 instances in ICE-TRAD

Feature	β	av. β	SD	p	av. p	SD
(Intercept)	1.34	1.30	0.17	0.033	0.056	0.042
PrnTh = p	1.47	1.46	0.24	0.002	0.007	0.007
GivTh = non	−1.97	−1.98	0.15	0.000	0.000	0.000
LenDif	−2.11	−2.12	0.04	0.000	0.000	0.000
AnRec = in	0.77	0.77	0.16	0.037	0.068	0.055
PrsRec = non	2.24	2.25	0.14	0.000	0.000	0.000

The coefficients β for the model fit are provided, together with the average βs in the ten separate models in the tenfold cross-validation, and their standard deviation. Also, the p-values for the model fit are shown, as well as the average p-values in the ten separate models, with their standard deviation

The results in Table 7 give an indication of the quality of the regression models. For the qualitative evaluation, we inspect the significant effects in the two models, shown in Tables 8 and 9.

Four of the five significant features in the traditional model are also found to be significant by the automatic model (printed above the horizontal line in Table 9). The effect that is missing in the automatic model is the pronominality of the theme. Instead, we have a significant effect for the pronominality of the recipient. For three features that are significant in both models (givenness of the theme, length difference and person of recipient), the signs of the β-coefficients are the same. This shows that the features have similar effects in both models. The exception is the animacy of the recipient, for which the sign is different in the two models. However, there are indications that both in the traditional model and in the automatic model, the effect is not very stable. First of all, the significance varies across the tenfolds: the average p-value is above 0.06, and the standard deviation is above 0.04. Second, we see a significant interaction of animacy with medium in the automatic model, in which the coefficient has the same direction as in the traditional model. Third, in a model that we built on ICE-AUTO without any interactions, the animacy of the recipient looses significance completely ($p > 0.90$). It also misses significance ($p < 0.10$) in a main-effects only model built on ICE-TRAD.

The automatic model has five additional significant effects (and a non-significant effect for medium that we kept because of the interactions), presented below the horizontal line. The definiteness of the theme is not significant across the ten folds, but its interaction with medium is. Also significant are the interaction of medium

Table 9 Significant features in the model built on the 889 instances in ICE-AUTO

Feature	β	av. β	SD	p	av. p	SD
(Intercept)	2.17	2.16	0.16	0.000	0.000	0.000
GivTh = non	−1.55	−1.55	0.13	0.000	0.001	0.000
LenDif	−1.80	−1.81	0.06	0.000	0.000	0.000
AnRec = in	−0.77	−0.77	0.15	0.038	0.063	0.046
PrsRec = non	1.10	1.11	0.08	0.003	0.005	0.003
ConTh = in	−1.46	−1.48	0.09	0.000	0.000	0.000
DefTh = in	0.86	0.86	0.16	0.039	0.062	0.046
PrnRec = p	−1.18	−1.19	0.14	0.000	0.001	0.001
Medium = W	0.16	0.16	0.16	0.711	0.691	0.200
DefTh = in, Medium = W	−1.48	−1.47	0.15	0.006	0.012	0.010
AnRec = in, Medium = W	1.33	1.33	0.20	0.011	0.022	0.019

Again, the coefficients β for the model fit are provided, together with the average βs in the ten separate models in the tenfold cross-validation, and their standard deviation. Also, the p-values for the model fit are shown, as well as the average p-values in the ten separate models, with their standard deviation

with the animacy of the recipient, the concreteness of the theme, and the pronominality of the recipient. Three of the five additional features thus involve the medium. In Sect. 3 we saw that the FDG parser has more problems with spoken data than with written data (resulting in a much lower precision). Now we see that this has substantially affected the regression model. Apparently, ICE-AUTO differs so much from the ICE-TRAD that it results in a qualitatively different model.

There are three ways to diminish the discrepancy between ICE-TRAD and ICE-AUTO: (1) by improving the precision of the detection of the cases, (2) by improving the recall of the detection of the cases, and (3) by improving the accuracy of the feature extraction. The second option would mean we either have to use a different parser, or we would have to extend the searches in the FDG parses. We believe this is beyond the scope of this article, and we will address this point in our general discussion in Sect. 6. The third option seems inefficient, since the accuracies reached by the feature extraction algorithm are already so high that they are surely very difficult to improve (cf. Table 6). We therefore choose to improve our data set with the first option: we improve the procedure by inserting a manual step between the detection of the candidates and the feature extraction, in which we manually filter the candidates found.[22] The result is the set of 633 instances, automatically annotated for the features (from now on referred to as ICE-SEMI). There is no significant difference between the proportion of instances from spoken material in ICE-SEMI (60.8%, 385/633) and in ICE-TRAD (60.0%, $\chi^2 = 0.07$, df = 1, $p > 0.75$). This is not true for the proportion of double object constructions: For ICE-SEMI, it is 79.1% (501/633), which is significantly different from the 72.3% in ICE-TRAD ($\chi^2 = 9.18$, df = 1, $p < 0.01$). But since the feature effects in logistic regression are

[22] All instances in the traditional set have already been checked manually by Theijssen (2010). In practice, we thus checked only the candidates that were not part of the traditional set.

Table 10 Significant features in the model built on the 633 instances in Ice-Semi

Feature	β	av. β	SD	p	av. p	SD
(Intercept)	1.85	1.85	0.40	0.018	0.040	0.030
PrnTh = p	1.17	1.18	0.27	0.040	0.070	0.050
GivTh = non	−2.20	−2.22	0.25	0.000	0.000	0.000
LenDif	−2.64	−2.66	0.12	0.000	0.000	0.000
PrsRec = non	1.23	1.24	0.24	0.015	0.028	0.027
PrnRec = p	−1.56	−1.57	0.20	0.001	0.000	0.000

Again, the coefficients β for the model fit are provided, together with the average βs in the ten separate models in the tenfold cross-validation, and their standard deviation. Also, the p-values for the model fit are shown, as well as the average p-values in the ten separate models, with their standard deviation

very robust against (increasing) class imbalance (Owen 2007), this will not influence our models.

For Ice-Semi, the model we found was very similar to the traditional model, with one main difference: the concreteness of the theme. It is highly significant in the semi-automatic model, while it did not come even near significance in the traditional model. After all our efforts in developing algorithms to establish concreteness automatically (see also Theijssen et al. 2011), we thus have to conclude that concreteness is too dependent on the context and on world knowledge to establish it automatically. For this reason, we decided to leave it out, and build a model with 23 features instead, i.e. all features we used before except the concreteness of the theme and its interaction with medium. The resulting model is the model presented in Tables 7 and 10.

We see that the effects found are indeed very similar to the ones for Ice-Trad in Table 8; the correlation between the five βs that are overlapping (those for Intercept, PrnTh, GivTh, LenDif and PrsRec) is 0.97.[23] In comparison with Ice-Auto, the rather unstable effect for the animacy of the recipient has now dropped out of significance, and the pronominality of the theme has become significant. In the comparison between Ice-Trad and Ice-Auto, we saw that the model built on Ice-Auto contained five significant effects more than the model built on Ice-Trad. When comparing the model built on Ice-Semi (Table 10) to the one built on Ice-Auto (Table 9), we see that three of these have now neatly disappeared: the interaction of medium with the animacy of the recipient, the definiteness of the theme and its interaction with medium.

The only extra effect that we have in comparison with the Ice-Trad model is that for the pronominality of the recipient. If we look at the distribution of this feature in the two data sets, we see that the proportion of pronominal recipients is higher in Ice-Semi (75.7%) than in Ice-Trad (65.6%). For both data sets, pronominal recipients occur more frequently in double object constructions than in prepositional dative constructions: 88.9% is in a double object construction in Ice-Semi, and

[23] The β for LenDif was first standardised by multiplying it by the standard deviation of LenDif in the data set.

87.0% in Ice-Trad. For non-pronominal recipients, there is no clear preference: 48.7% are in a double object constructions in Ice-Semi, and 44.1% in Ice-Trad. The pronominality of the recipient thus seems to have a similar distribution with respect to the dative alternation in the two data sets. It only shows up as significant in the Ice-Semi model because pronominal recipients form a bigger proportion of that set.

We thus conclude that once the low precision of the automatic instance detection is cured, and the concreteness of the theme is left out of consideration, the model is very similar to what we find with a data set that was established completely manually. The semi-automatic model is not really affected by the recall of the detection or the smaller size of the data set. Although we aimed for a completely automatic approach, we have to conclude that human intervention is required, at least when using an off-the-shelf parser like the FDG parser we employed.

5.3 Results for the Switchboard data

In this section, we perform an extrinsic evaluation on the test data, the Switchboard data. Given the conclusions of the previous section, we compare the following two models: (1) a model built on the 2,349 instances in Swb-Trad, and (2) a model built on the semi-automatic set with the 1,292 instances that were also found automatically (Swb-Semi). In Swb-Semi, the proportion of double object construction is 83.0% (1,073/1,292), being significantly higher than the proportion of 78.8% in Swb-Trad ($\chi^2 = 9.43$, df = 1, $p < 0.01$). Again, this is not a problem because logistic regression is robust against class imbalance.

The concreteness of the theme was again excluded from the feature set. The quality of the models is summarised in Table 11. Both models show a very good fit to the data, with hardly any overfitting.

The significant effects in the regression models can be found in Tables 12 and 13. Both show significant effects for the definiteness of the recipient and the theme, the givenness of the theme and the length difference between the theme and the recipient. The coefficients also show the same polarity, and their correlation is high (0.97).[24]

The semi-automatic model contains one extra effect: the number of the recipient. As we found for the person of the recipient in the intrinsic evaluation (Sect. 4), this difference seems to be the result of slight differences in the annotation guidelines. Whereas we use a purely syntactic definition, the annotation in Bresnan et al. (2007) is semantic. For instance, when speaking about a hypothetical person, speakers sometimes switch to plural *them* to refer to such persons. We label *them* as plural, while the annotations in Swb-Trad call it singular. This was the case in 38 of the 64 disagreeing annotations. For 14 more, the disagreement was caused by a different treatment of the noun *people*, being semantically plural (*a group of persons*), but syntactically singular (plural: *peoples*).

The traditional model contains four more significant effects that were not found in the semi-automatic model: the pronominality of the theme, the animacy of the recipient, the givenness of the recipient and the pronominality of the recipient.

[24] Again, we standardised length difference by multiplying the β by the standard deviation of the feature in the data.

Table 11 Prediction accuracy and concordance C for the model fit (*Acc, C*) and in tenfold cross-validation (*av. Acc, av. C*), for SWB-TRAD and SWB-SEMI

Data set	Majority	N	Acc	av. Acc	SD	C	av. C	SD
SWB-TRAD	0.788	2,349	0.933	0.927	0.015	0.972	0.967	0.014
SWB-SEMI	0.830	1,292	0.957	0.954	0.026	0.975	0.969	0.021

The majority baseline and the number of instances are also provided

Table 12 Significant features in the model built on the 2,349 instances in SWB-TRAD

Feature	β	av. β	SD	p	av. p	SD
(Intercept)	0.30	0.32	0.14	0.605	0.602	0.153
DefRec = in	0.89	0.89	0.09	0.003	0.007	0.008
DefTh = in	−1.62	−1.63	0.13	0.000	0.000	0.000
PrnRec = p	−0.78	−0.78	0.09	0.008	0.015	0.011
PrnTh = p	1.49	1.48	0.08	0.000	0.000	0.000
GivTh = non	−1.43	−1.43	0.11	0.000	0.000	0.000
GivRec = non	1.31	1.32	0.08	0.000	0.000	0.000
LenDif	−1.61	−1.62	0.07	0.000	0.000	0.000
AnRec = in	1.87	1.88	0.07	0.000	0.000	0.000

The coefficients β for the model fit are provided, together with the average βs in the ten separate models in the tenfold cross-validation, and their standard deviation. Also, the p-values for the model fit are shown, as well as the average p-values in the ten separate models, with their standard deviation

Table 13 Significant features in the model built on the 1,292 instances in SWB-SEMI

Feature	β	av. β	st.dev	p	av. p	st.dev.
(Intercept)	1.54	1.57	0.15	0.009	0.012	0.008
DefRec = in	3.55	3.56	0.20	0.000	0.000	0.000
DefTh = in	−2.09	−2.10	0.20	0.000	0.001	0.002
GivTh = non	−1.58	−1.59	0.20	0.004	0.007	0.005
LenDif	−3.60	−3.62	0.16	0.000	0.000	0.000
NrRec = sg	−0.93	−0.93	0.06	0.003	0.005	0.002

Again, the coefficients β for the model fit are provided, together with the average βs in the ten separate models in the tenfold cross-validation, and their standard deviation. Also, the p-values for the model fit are shown, as well as the average p-values in the ten separate models, with their standard deviation

For the latter three, the problem is that they are correlated: all three very frequently have the same value. This is because many recipients consist of personal pronouns only, and they are always pronominal, definite and discourse given, and animate most of the time. Because these features are correlated, it is not surprising that not all of them show up in the semi-automatic model, which is based on fewer data points than the traditional model.

As with the models for the ICE data, we see some differences between the models built on Swb-Trad and on Swb-Semi. But these differences do not seem to be caused by the quality of the automatic approach, but by difficulties in the data itself: the use of different annotation definitions and the correlation of many of the features. The low recall of 55.0% may explain the lack of significance for some of those correlated features.

6 General discussion

Besides the difficulty of collecting a suitable data set that can be used to model variation in language (e.g. syntactic alternation), linguists taking such a modelling approach have a more fundamental challenge to meet. When using modelling techniques such as logistic regression, one models the data that is offered. Two different data sets, though drawn from the same population, can result in different models because their composition differs. Because we use two different samples (traditional and automatic) from the English language as represented in the ICE-GB and Switchboard corpora, this accidental composition could affect the models. It is not clear whether the traditional set is closer to the actual English language than the automatic set. The models found for either of the data sets are not necessarily true, and the features that show no significance in our models could still play a role in another data set. Moreover, there is still no consensus about the definitions of the features we have employed. The definitions we used for this article are chosen such that they allow comparison with previous work (Bresnan et al. 2007; Theijssen 2010), but they are by no means definitive. Moreover, we have seen that even the definitions of which we believed they were the same, appeared to be slightly different after all.

The effect of the composition of a data set usually grows when data sets become smaller. In the near future, we will therefore apply the procedure to a larger corpus: the one-hundred-million-word British National Corpus (BNC Consortium, 2007). The results found for this data set may show whether the almost significant effects turn up really significant when larger amounts of data are considered. A possible drawback is that we have shown that a fully automatic approach is not accurate enough. Instead, we need a semi-automatic approach in which we manually filter the candidates suggested by the parser. For a large corpus as the BNC, this step may take considerable time. However, some preliminary annotation work shows that the manual checking is not as time-consuming as one would think: with the help of a user-friendly interface, one can check up to 200 candidates per hour. Moreover, the inter-annotator agreement for this task is comforting: an average κ of 0.74 (for four annotators who all checked the same 100 candidates).

One might wonder if the human intervention is still needed when employing a different parser. In this article, we have decided to use an off-the-shelf syntactic parser that distinguishes both dative constructions explicitly. Parsers that have this information available are rare, and we believe human intervention will always be necessary. Of course, such a manual step, in which one checks the candidates suggested by the parser, can also be performed on the output of other parsers that may or may not recognise dative constructions explicitly. One could for instance decide to employ a parser that does not distinguish between prepositional dative constructions and locative constructions (e.g. *I brought him to school*), but that yields a higher recall. Another possibility is to improve an existing parser by training on data that is similar to the data studied. However, this is a difficult procedure that requires one to have experience with parsing. Our approach has shown that even with an off-the-shelf parser, that yields a low recall, sensible data sets can be obtained. This is a promising result for corpus linguists who study a syntactic phenomenon but do not have access to syntactically annotated data.

In fact, the semi-automatic approach is also suitable for research on different syntactic alternations. One could select a parser that seems to perform well on the construction in question, and then manually check the proposed candidates. When seeing recurring patterns, one can add simple heuristic rules like we formulated for the dative alternation. Next, one can use the feature extraction script presented in this article.[25] Many of the features included in the script are generally known to be relevant for other syntactic alternations, as already noted in Sect. 1. The script should be provided with three bits of information for each noun phrase that needs annotation: (1) which word is the syntactic head, (2) what are the lemmas of the words in the noun phrase, and (3) what are the POS tags of these words. Using a different parser would thus mean that the extraction script needs some adjustments. For establishing the discourse givenness, it also needs to have the preceding context. The selection of a corpus thus also leads to the need for some minor changes in the extraction script, so it can deal with the corpus input provided.

7 Conclusion

In this article, we have addressed the question of whether automatically obtained and enriched data is suitable for use in linguistic research on syntactic alternations, even if the data may contain errors. We have taken the English dative alternation as a case study. This offered us a way to evaluate the automatically obtained data extrinsically, namely by employing it to build logistic regression models like those in Bresnan et al. (2007). We employed two data sets that were manually obtained: 930 instances collected by Theijssen (2010) from the ICE-GB corpus of spoken and written British English (Ice-Trad), and 2,349 instances collected by Bresnan et al. (2007) from the Switchboard corpus of spoken American English (Swb-Trad). The first data set has been employed to tailor the automatic approaches, and to evaluate the errors made. The second data set has not been seen previously, and has been used as a test set in quantitative evaluations. With respect to the aforementioned question, there are two main conclusions to be drawn.

First, we have to conclude that the FDG parser that we employed is not very successful in detecting instances of the dative alternation. In combination with our filtering heuristics, the recall was 66.6% for the instances found automatically in the ICE-GB (Ice-Auto) and 55.0% for those found in Switchboard (Swb-Auto). For precision, we reached 69.6% for Ice-Auto and only 48.0% for Swb-Auto. The

[25] The feature extraction script can be downloaded from http://lands.let.ru.nl/~daphne/downloads.

analysis of the errors in Ice-Auto showed that the FDG parser has most difficulty with spoken material, with longer sentences and with PP-attachment. Parse errors were the main cause of missing instances (decreasing recall) and incorrectly accepting candidates (decreasing precision). Seeing the nature of the Switchboard data (spontaneous speech only, with many disfluencies), it is not surprising that the FDG parser has great difficulty recognising dative constructions. The regression model for Ice-Auto contained four significant effects that were not found for Ice-Trad. We concluded that Ice-Auto contained too many errors to give the same—or at least similar—results as those obtained for Ice-Trad. We solved this problem by inserting one (simple) manual step: manually checking the relevance of the candidates that were found automatically, before annotating the approved instances automatically. The model built on only the 633 instances that were manually approved (Ice-Semi) appeared to be very similar to the one found for Ice-Trad. This is also what we found for the 1,292 approved candidates in the semi-automatic Switchboard set (Swb-Semi).

Second, we conclude that our rather straight-forward feature extraction algorithm is suitable for automatically annotating the instances with information that is syntactic (e.g. number), semantic (e.g. animacy) and discourse-related (e.g. givenness) in nature. The κ scores between the manual and the automatic annotations were similar to scores found between human annotators, except for the intuitively most difficult features: animacy, concreteness and discourse givenness. Only the automatic annotation of the concreteness of theme was so dissimilar from the human annotations that it notably influenced the regression models. When excluding this feature, the models built on Ice-Semi and Swb-Semi (with the automatic annotations) were very similar to the ones obtained for Ice-Trad and Swb-Trad (with manual annotations). The differences we found did not seem to be caused by the errors in the automatic annotations, but by properties inherent to the data sets: multiple correlations between the features, and the presence of different definitions for the same feature.

In sum, we see that the models found for the automatic data sets are especially hampered by the presence of candidates that are not really instances of the dative alternation, but that were included due to errors in the automatic analyses. We also have to conclude that establishing the concreteness of nouns automatically is a bridge too far. But when the instances found are manually checked for relevance, and concreteness is left out of consideration, the models found are very similar to the ones found for traditionally established data sets.

Appendix: annotation of the features in the traditional sets

Animacy of recipient (AnRec)

Following Bresnan et al. (2007), the animacy of the recipient was annotated as a binary feature: it was labelled either *animate* (human and animal) or *inanimate* (not human or animal). Companies and organizations were considered animate when it was evident from the context that the writer meant the people working in these institutions.

Concreteness of theme (ConTh)

For the annotation of the concreteness of the theme, the instructions in Bresnan et al. (2007) were not very clear, except that the feature again allowed only two values: either *concrete* or *abstract*. We decided to follow Garretson (2003), in which a noun phrase is deemed concrete if it is prototypically concrete. We assumed that prototypically concrete objects have a known physical size. The themes that did not fit this description were labelled *abstract*.

Definiteness of recipient and theme (DefRec, DefTh)

For both the recipient and the theme we annotated the definiteness. All (syntactic) object heads that were preceded by a definite article, a genitive form or a definite pronoun (e.g. demonstrative and possessive pronouns), and all objects that were proper nouns or definite pronouns themselves, were annotated as *definite*. The remaining objects were given the value *indefinite*.

Discourse givenness of recipient and theme (GivRec, GivTh)

A recipient or theme was labelled *given* when it was mentioned in the preceding context (maximally 20 clauses before). We also considered an object given when it was stereotypical of something mentioned in the preceding context, or when it was part of the writing context (e.g. the newspaper article itself, or the reader). *You, one* and *us* as impersonal pronouns were annotated as given as well. All remaining objects received the value *new*.

Number of recipient and theme (NrRec, NrTh)

Recipients and themes were annotated for number: *singular* or *plural*. In case a recipient or theme could refer to something singular or plural (which is especially the case with the pronoun *you*), the antecedent was checked.

Person of recipient (PrsRec)

Person of recipient was annotated by giving it the value *local* or *nonlocal*. Local recipients are in first or second person (e.g. *I, me, yourself*), non-local ones in third person.

Pronominality of recipient and theme (PrnRec, PrnTh)

We also annotated whether the recipient and the theme were (syntactically) headed by a pronoun and thus *pronominal*, or not (*nonpronominal*). We treated all types of pronouns as such, including for instance indefinite and relative pronouns like *all* and *that*.

Length difference (LenDif)

An important factor in clause word order is the so-called *principle of end weight* (e.g. Quirk et al. 1972), which states that language users tend to place the more complex constituents at the end of an utterance. Bresnan et al. (2007) therefore included a feature indicating the length difference between the recipient and the theme. Following their approach, we used a Perl script that counts the number of words in the recipient and the theme by splitting on white space, and takes the natural log of these lengths to smoothen outliers. The length difference is then calculated by subtracting the recipient length from the theme length.

References

Agirre, E., Baldwin, T., & Martinez, D. (2008). Improving parsing and PP attachment performance with sense information. In *Proceedings of the workshop on human language technologies at the 46th annual meeting of the association for computational linguistics (ACL-08)* (pp. 317–325).

Anagnostopoulou, E. (2005). Cross-linguistic and cross-categorial variation of datives. In M. Stavrou & A. Terzi (Eds.), *Advances in Greek generative grammar* (pp. 61–126). Amsterdam: John Benjamins.

Baker, K., & Brew, C. (2008). *Multilingual animacy classification by sparse logistic regression*. Ohio State working papers in linguistics.

Bean, D. L., & Riloff, E. (1999). Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th annual meeting of the association for computational linguistics (ACL'99)* (pp. 373–380).

Beavers, J., & Nishida, C. (2010). The Spanish dative alternation revisited. In S. Colina, A. Olarrea, & A. Carvalho (Eds.), *Romance linguistics 2009: Selected papers from the 39th linguistic symposium of romance languages* (pp. 217–230).

Bikel, D. M. (2002). Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of the human language technology conference 2002 (HLT-2002)*.

Blackwell, A. A. (2005). Acquiring the English adjective lexicon: relationships with input properties and adjectival semantic typology. *Child Language, 32*(3), 535–562.

BNC Consortium. (2007). *The British national corpus, version 3* (BNC XML Edition). Oxford University Computing Services, http://www.natcorp.ox.ac.uk.

Bod, R. (2007). Is the end of supervised parsing in sight? In *Proceedings of the 45th annual meeting of the association for computational linguistics (ACL 2007)* (pp. 400–407).

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.

Chang, C. C., & Lin, C. J. (2001). *LIBSVM: A library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/cjlin/libsv.

Changizi, M. A. (2008). Economically organized hierarchies in WordNet and the Oxford English Dictionary. *Cognitive Systems Research, 9*(3), 214–228.

Colleman, T. (2006). *De Nederlandse datiefalternantie: een constructioneel en corpusgebaseerd onderzoek [The dative alternation in Dutch: a constructional and corpus-based study]*. PhD thesis, Ghent University.

Collins, P. (1995). The indirect object construction in English: An informational approach. *Linguistics, 33*(1), 35–49.

Coltheart, M. (1981). *MRC psycholinguistic database user manual: Version 1*. Birkbeck College.

Daudaravičius, V., & Marcinkevičiene, R. (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics, 9*(2), 321–348.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, Massachusetts, USA: MIT Press.

Garretson, G. (2003). *Coding manual for the project "Optimal typology of determiner phrases"*. Unpublished manuscript, Boston University.

Girju, R., Roth, D., & Sammons, M. (2005). Token-level disambiguation of VerbNet classes. In K. Erk, A. Melinger, & S. Schulte im Walde (Eds.), *Proceedings of the interdisciplinary workshop on the identification and representation of verb features and verb classes* (pp. 56–61).

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92* (pp. 517–520) San Fransisco, USA.

Gomes, C. A. (2003). Dative alternation in Brazilian Portuguese: Typology and constraints. *Language Design: Journal of Theoretical and Experimental Linguistics, 5*, 67–78.

Greenbaum, S. (Ed.) (1996). *Comparing English worldwide: The international corpus of English*. Oxford, UK: Clarendon Press.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.) *A mosaic of corpus linguistics: Selected approaches* (pp. 269–291). Germany: Peter Lang, Frankfurt am Main.

Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'Alternations'. *International Journal of Corpus Linguistics, 9*(1), 97–129.

Grimm, S., & Bresnan, J. (2009). *Spatiotemporal variation in the dative alternation: A study of four corpora of British and American English*. Third International Conference Grammar and Corpora.

Grondelaers, S., & Speelman, D. (2007). A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory, 3*(2), 161–193.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations, 11*(1), 10–18.

Harris, Z. (1954). Distributional structure. *Word, 10*(23), 146–162.

Hinrichs, L., Smith, N., & Waibel, B. (2007). *The part-of-speech-tagged 'Brown' corpora: A manual of information, including pointers for successful use*. Department of English, Albert-Ludwigs-Universität Freiburg.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning*. New York, USA.: Springer

Joanis, E., Stevenson, S., & James, D. (2008). A general feature space for automatic verb classification. *Natural Language Engineering, 14*(3), 337–367.

Kabadjov, M. A. (2007). *A comprehensive evaluation of anaphora resolution and discourse-new classification*. PhD thesis. Department of Computer Science, University of Essex.

Keller, F., Corley, M., Corley, S., Crocker, M. W., & Trewin, S. (1999). Gsearch: A tool for syntactic investigation of unparsed corpora. In *Proceedings of the EACL workshop on linguistically interpreted corpora* (pp. 56–63).

Kipper, K., Dang, H. T., & Palmer, M. (2000). Class-based construction of a verb lexicon. In *Proceedings of the 17th national conference on artificial intelligence (AAAI-2000)* (pp. 691–696).

Korhonen, A. (2009). Automatic lexical classification—Balancing between machine learning and linguistics. In *Proceedings of the 23rd Pacific Asia conference on language, information and computation* (pp. 19–28).

Kuijjer, C. (2007). *Semantic lexicon expansion using bootstrapping and syntax-based, contextual extraction patterns*. Master's thesis, Information Sciences, University of Amsterdam.

Lapata, M. (1999). Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th annual meeting of the association for computational linguistics (ACL'99)* (pp. 397–404).

Lapata, M., & Brew, C. (2004). Verb class disambiguation using informative priors. *Computational Linguistics, 30*(1), 45–73.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago, USA: The University of Chicago.

Li, J., & Brew, C. (2008). Which are the best features for automatic verb classification. In *Proceedings of the 46th annual meeting of the association for computational linguistics* (pp. 434–442).

McCarthy, D. (2001). Lexical acquisition at the syntax-semantics interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. PhD thesis, University of Sussex.

Nancarrow, O., & Atwell, E. (2007). A comparative study of the tagging of adverbs in modern English corpora. In *Proceedings of corpus linguistics 2007 (CL2007).*

Ng, V., & Cardie, C. (2002). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th international conference on computational linguistics (COLING-2002)* (pp. 730–736).

Oostdijk N. (1996) Using the TOSCA analysis system to analyse a software manual corpus. In R. Sutcliffe, H. Koch, & A. McElligott (Eds.), *Industrial parsing of software manuals*. Amsterdam: The Netherlands: Rodopi.

Orăsan, C., & Evans, R. (2007). NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research, 29*, 79–103.

Owen, A. B. (2007). Infinitely imbalanced logistic regression. *The Journal of Machine Learning Research, 8*, 761–773.

Poesio, M., Uryupina, O., Vieira, R., Kabadjov, M. A., & Goulart, R. (2004). Discourse-new detectors for definite description resolution: A survey and a preliminary proposal. In *Proceedings of the workshop on reference resolution at the 42nd annual meeting of the association for computational linguistics (ACL-04)* (pp. 47–54).

Poibeau, T., & Messiant, C. (2008). Do we still need gold standards for evaluation? In *Proceedings of the language resources and evaluation conference (LREC)* (pp. 547–552).

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1972). *A grammar of contemporary English*. London, UK: Longman.

R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Riloff, E., & Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th national conference on artificial intelligence (AAAI-99)* (pp. 474–479).

Rooth, M., Riezler, S., Prescher, D., Carroll, G., & Beil, F. (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th annual meeting of the association for computational linguistics (ACL'99)* (pp 104–111).

Rosenbach, A. (2003). Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In G. Rohdenburg & B. Mondorf (Eds.), *Determinants of grammatical variation in English*. Berlin, Germany: De Gruyter

Schulte im Walde, S. (2009). The Induction of verb frames and verb classes from corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (Vol. 2, Chap. 44, pp. 952–972). Berlin, Germany: Mouton de Gruyter.

Schulte im Walde, S., Hying, C., Scheible, C., & Schmid, H. (2008). Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In *Proceedings of the 46th annual meeting of the association for computational linguistics* (pp. 496–504).

Sun, L., & Korhonen, A. (2009). Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 conference on empirical methods in natural language processing (EMNLP 2009)* (pp. 638–647).

Tapanainen, P., & Järvinen, T. (1997). A non-projective dependency parser. In *Proceedings of the 5th conference on applied natural language processing (ANLP)* (pp. 64–71).

Theijssen, D. (2010). Variable selection in logistic regression: The British English dative alternation. In T. Icard & R. Muskens (Eds.), *Interfaces: Explorations in logic, language and computation, Springer, lecture notes in computer science (subseries: Lecture notes in artificial intelligence)* (Vol. 6211, pp. 87–101).

Theijssen, D., Verberne, S., Oostdijk, N., & Boves, L. (2007). Evaluating deep syntactic parsing: Using TOSCA for the analysis of *why*-questions. In P. Dirix, I. Schuurman, V. Vandeghinste, & F. Van Eynde (Eds.), *Computational linguistics in the Netherlands 2006: Selected papers from the seventeenth CLIN meeting, no. 7 in LOT Occasional Series* (pp. 115–130).

Theijssen, D., van Halteren, H., Boves, L., & Oostdijk, N. (2011). On the difficulty of making concreteness concrete. *Computational Linguistics in the Netherlands (CLIN 21)*.

Thelen, M., & Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002)* (pp. 214–221).

Uryupina, O. (2003). High-precision identification of discourse new and unique noun phrases. In *Proceedings of the student workshop at the 41st annual meeting of the association for computational linguistics (ACL-03)* (pp. 80–86).

Vieira, R. (1998). *Definite description resolution in unrestricted texts*. PhD thesis, Centre for Cognitive Science, University of Edinburgh.

Vieira, R., & Poesio, M. (2000). An empirically based system for processing definite descriptions. *Computational Linguistics, 26*(4), 539–593.

Xing, X., Zhang, Y., & Han, M. (2010). Query difficulty prediction for contextual image retrieval. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke et al. (Eds.), *Proceedings of the 32nd European Conference on Information Retrieval (ECIR 2010)* (pp. 581–585).