Modeling Dative Alternations of Individual Children

Antal van den Bosch*, Joan Bresnan**
*Centre for Language Studies, Radboud University Nijmegen, The Netherlands
**Center for the Study of Language and Information, Stanford University, Stanford, CA, USA

Abstract. De Marneffe et al. (2012) model the production of the dative alternation in English by seven young children, using data from the Child Language Data Exchange System corpus. Using mixed-effects logistic modelling on the aggregated data of these children, De Marneffe *et al.* report that the children's choices can be predicted both by their own utterances and by child-directed speech. We present a follow-up study that brings the computational modeling down to the individual child, using memory-based learning and incremental learning curve studies. We observe that for all children, their dative choices are best predicted by a model trained on child-directed speech. Yet, models trained on two individual children for which sufficient data is available are about as accurate. Furthermore, models trained on the dative alternations of these children provide approximations of dative alternations in caregiver speech that are about as accurate as training and testing on caregiver data only.

1 Introduction

The production of language is the result of a great number of choices made by the individual speaker, where each choice may be affected by various factors that, according to a large body of work, range from simple word frequencies to subtle semantic factors. For instance, which variant of the dative alternation speakers produce has been shown in a corpus study to be partially affected by the animacy and givenness of the recipient and theme (Bresnan et al. 2007). An inanimate recipient tends to co-occur with a prepositional dative construction ("bring more jobs and more federal spending to their little area").

Somehow and at some point in language acquisition, children learn these preferences, but it takes several years before children approximate adult language use. Monitoring and modeling this process of development may shed light on the inner workings of language learning in general, but to keep experiments under control, most studies, including the one presented here, zoom in on a representative but specific phenomenon. This contribution continues a line of research introduced by De Marneffe et al. (2012), who formulate three general research questions: (1) do children show sensitivity to linguistic probability in their own syntactic choices, and if so, (2) are those probabilities driven by the same factors that affect adult production?

And finally, (3) do children assign the same weight to various factors as their caretakers? If so, then this may support the hypothesis that from early on children are sensitive to complex distributional patterns.

Syntactic alternations such as the genitive, dative, or locative alternation in English are choices that speakers have in generating different syntactic forms that carry approximately the same meaning. Monitoring speakers and observing which particular choices they make in which context allows us to explore the predictive components in this context from which we can guess which choice is going to be made.

The English dative alternation, the focus of this contribution, refers to the choice between a prepositional dative construction (NP PP) as in "I gave the duck to my Mommy", where the NP is the theme and the PP contains the recipient, and a double object construction (NP NP) as in "I gave my Mommy the duck", where the first NP is the recipient and the second NP is the theme. A robust finding accross studies is that inanimate, indefinite, nominal, or longer arguments tend to be placed in the final complement position of the dative construction, while animate, definite, pronominal, or shorter arguments are placed next to the verb, preceding the other complement (De Marneffe et al. 2012). This means, for instance, that if a recipient of the dative construction is pronominal, such as *me*, it will tend to occur immediately after the verb, triggering a double object dative.

The dative construction is frequently used by children as well as their caregivers in child-directed speech (Campbell and Tomasello 2001); this makes it a suitable focus for modeling syntactic alternations in child production.

While De Marneffe et al. (2012) use mixed-effects logistic regression to model dative alternation in children's speech, Theijssen (2012) compares regression-based and memory-based learning accounts of the dative alternation choice in adults. Theijssen's dataset consisted of 11,784 adult constructions of both types extracted from the British National Corpus (Burnard 2000), 7,757 of which occur in transcribed spoken utterances, 4,027 in written sentences. Her mixed-effects logistic regression approach uses automatically extracted higher-level determinants: animacy, definiteness, givenness, pronominality, and person of the recipient, and definiteness, givenness, and pronominality of the theme. Alternatively, Theijssen applied a memory-based learning classifier (Daelemans and Van den Bosch 2005) which we also apply in this study. The memory-based approach she used included lexical information only: the identity (stem) of the verb, the recipient, and the theme.

Theijssen reports that MBL classifies unseen cases about as accurately (93.1% correct) into the two dative choices as regression analysis does, which attains a fit of 93.5%, while MBL does so without the higher-level features. According to Theijssen, the main factors for the success of the simplistic MBL approach are the strong licensing of one or the other dative construction by particular verbs, and the significant effect of length difference between recipient and theme. Both aspects of the input can be learned directly from from lexical input, while they remain hidden in

the higher-level features. In this study we keep the available features identical to the earlier approach introduced by De Marneffe et al. (2012) in order to stay close to this particular study, which focused on datives with two verbs only (*give* and *show*).

2 Modeling learning curves of individual children

2.1 Memory-based learning

Memory-based learning is a computational approach to solving natural language processing problems. The approach is based on the combination of a memory component and a processing component. Learning happens by storing attested examples of the problem in memory. New unseen examples of the same problem are solved through similarity-based reasoning on the basis of the stored examples (Daelemans and Van den Bosch 2005). In other words, memory-based learning offers a computational implementation of example-based or exemplar-based language processing.

Van den Bosch and Daelemans (2013) argue that from a cognitive perspective the approach is attractive as a model for human language processing because it does not make any assumptions about the way abstractions are shaped, nor does it make any a priori distinction between regular and exceptional exemplars, allowing it to explain fluidity of linguistic categories, and both regularization and irregularization in processing.

As a software tool for our experiments we use TiMBL[1] (Daelemans et al. 2010). In all our experiments we use the default setting of this implementation, which is based on the IB1 algorithm (Aha et al. 1991) and which adds an information-theoretic feature weighting metric. When the memory-based learning algorithm is asked to predict the class of an unseen test exemplar, it compares it to all training exemplars in memory, and constructs a ranking of the k nearest (or most similar) neighbors. The class that the algorithm predicts for the new exemplar is the majority class found among the k nearest neighbors.

To compute the similarity between an unseen test exemplar and a single training exemplar, the Overlap similarity function is used, weighted by gain ratio (Daelemans et al. 2010), expressed in Equation 1.1:

$$\Delta(X,Y) = \sum_{i=1}^{n} w_i \, \delta(x_i, y_i) \tag{1.1}$$

where:

1. TiMBL, Tilburg Memory-Based Learner, is an open-source toolkit available from `http://ilk.uvt.nl/timbl`. We used version 6.4.5.

$$\delta(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases} \tag{1.2}$$

and w_i represents the gain-ratio weight of feature i:

$$w_i = \frac{H(C) - \sum_{v \in V_i} P(v) \times H(C|v)}{si(i)} \tag{1.3}$$

Where C is the set of class labels, $H(C) = -\sum_{c \in C} P(c) \log_2 P(c)$ is the entropy of the class labels, V_i is the set of values for feature i, and $H(C|v)$ is the conditional entropy of the subset of the training examples that have value v on feature i. The probabilities are estimated from relative frequencies in the training set. Finally, $si(i)$ is the so-called split info, or the entropy of the values, of feature i (Quinlan 1993):

$$si(i) = -\sum_{v \in V_i} P(v) \log_2 P(v) \tag{1.4}$$

The gain ratio weighting assigns higher weights to features that are more predictive with respect to the class. It is more robust than the simpler information gain metric, which overestimates the importance of features with many values (such as lexical features); the split info, the entropy of the values, acts as a penalty for a feature with many values. One effect of this weighting in the similarity function is that mismatches on features with a large gain ratio cause memory exemplars to be more distant than when the mismatch is on features with a small gain ratio. On the other hand, the gain ratio weight of a feature may be so prominent that it promotes a memory exemplar with a matching value on that feature to the top-ranking k nearest neighbors, despite the fact that other less important features carry non-matching values.

Memory-based learning can be likened to local regression or locally-weighted learning (Atkeson et al. 1997). It has similar issues with feature collinearity (gain ratio weights are computed separately for each feature; redundancy is not taken into account), but by limiting its decision to local evidence found close to the test exemplar, the algorithm is sensitive to subtle co-occurences of matching features in the k nearest neighbors.

The default version of TiMBL, used in this study, sets the number of neighbors to $k = 1$, which implies that an unseen test vector is compared to all training exemplars, and the dative choice label of the single most similar training exemplar is taken as the prediction of the test exemplar.

3 Experimental setup

3.1 Data collection

We used the same data as De Marneffe et al. (2012), which were extracted from the Child Language Data Exchange System (CHILDES) (MacWhinney 2000), a publicly available database of children's speech produced in a natural environment. De Marneffe *et al*. focused on seven children: Abe, Adam, Naomi, Nina, Sarah, Shem, and Trevor, based on the amount of data available for them compared to other children, in terms of both their total number of utterances and the number of utterances containing one of the variants of the dative alternation. The utterances were taken from the children's production between the ages of 2–5 years. The data yielded a sufficient number of utterances to investigate two verbs in depth, *give* and *show*, which are the only ones considered in this study. On top of this filtering, De Marneffe *et al*. selected only dative constructions following the "verb NP NP" (double object construction) or "verb NP PP" (prepositional dative) pattern.

For all seven children, conversations with caregivers were included as well. Table 1 lists the basic statistics of available child and child-directed utterances with dative alternations, and the age range of the individual children (in days). For two children, Adam and Nina, we have more than one hundred dative attestations in their own speech. For both children we also have more than one hundred datives in the speech directed to them by their caregivers; for a third child, Shem, we also have over a hundred caregiver utterances containing datives.

Table 1. Basic statistics for the seven children used in the study: numbers of utterances and age range in days.

	# Datives in		Age (days) of attestation	
Child	child data	child-directed speech	First	Last
Abe	74	0	924	1,803
Adam	221	207	824	1,897
Naomi	21	0	767	1,733
Nina	146	443	747	1,193
Sarah	19	0	1,178	1,841
Shem	15	138	875	1,130
Trevor	33	0	757	1,452

Following the encoding of the data by De Marneffe *et al*. in their computational modeling experiment with mixed-effects logistic regression, all attestations of both dative constructions in their utterance context are converted to feature vectors. Each vector (examplar) is metadated with the exact day of attestation, and labeled with

the dative choice (i.e. a binary choice between the double object construction and the prepositional dative). Each vector is composed of fourteen feature values; the fourteen underlying features are listed in Table 2.

Table 2. The fourteen features used in the study, along with their gain ratio based on a concatenation of all children's data.

Name	Description	Gain ratio
Prime	The type of nearest previous occurrence of a dative construction, if any, within the 10 preceding lines. Three values are distinguished: 0 = none, NP = double NP-dative ("give me a hug"); PP = to-dative ("give it to me")	0.076
Verb	"give" or "show"; the two most frequent dative verbs collected in the childrens's speech	0.006
Theme	that which shown or given ("a hug" in "give me a hug"; "it" in "give it to me")	0.079
Recipient	to whom or which the theme is shown or given ("me" in "give me a hug" and "give it to me")	0.079
Theme givenness levels	either 'given': the referent of the theme was mentioned in the preceding ten lines or was denoted by a first or second person pronoun, "me", "us", or "you"; or 'new': not given	0.038
Recipient givenness levels	coded in the the same way as *Theme givenness levels*	0.086
Theme animacy	1 = the theme refers to a human or animal; 0 = other	0.022
Recipient animacy	coded in the same way as *Theme animacy*	0.005
Theme toy animacy	explicitly encodes toy themes as animate: 1 = the theme refers to a human or animal or toy; 0 = not animate	0.000
Recipient toy animacy	coded in the same way as *Theme toy animacy*	0.051
Theme pronoun status	'pronoun' = the theme is a definite pronoun ("it", "them") or a demonstrative pronoun ("this", "dis", "those", etc); 'lexical' = not pronoun	0.276
Recipient pronoun status	coded in the same way as *Theme pronoun status*	0.113
Theme corrected lenght	length of the theme in orthographic words	0.071
Recipient corrected length	length of the recipient in orthographic words	0.086

The Theme and Recipient length features are *corrected* due to the fact that in the original data used by De Marneffe *et al.* some recipients and themes mistakenly included other material such as adverbials.

The third column of Table 3 lists the gain ratio weights for each feature (cf. Equation 1.3). These weights seem to suggest four groups of features:

1. *Theme pronoun status* and *Recipient pronoun status* are by far the most predictive features. *Theme pronoun status* has a weight about 2.5 times higher than that of *Recipient pronoun status*, and over three times higher than the third highest weight;

2. There is a second-tier group of informative features with a gain ratio of about $0.07 - 0.08$: *Prime*, *Theme*, *Recipient*, *Recipient givenness levels*, *Theme corrected lenght*, and *Recipient corrected length*;

3. A third-tier group of features has weights in the range of $0.02 - 0.05$: *Theme givenness levels*, *Theme animacy*, and *Recipient toy animacy*;

4. A fourth-tier group has near-zero weights, carrying hardly any predictive information: *Verb*, *Recipient animacy*, and *Theme toy animacy*.

Perhaps somewhat surprisingly, the identity of the verb (*give* or *show*) is virtually unrelated to the dative choice. In other words, the identity of the verb does not license one of the dative constructions.[2] The high weights for the pronoun status features imply that the likelihood of being a nearest neighbor is large when it has the same values on either of these features as the test exemplar. Yet, the weights of the other features, especially those in the second-tier group, are large enough to outweigh a mismatch on the pronoun features.

3.2 Learning curve evaluation

Our experiments are run per individual child, in an iterative experiment that tracks the child on a day-by-day basis and computes a learning curve. Figure 1 illustrates how the iterative learning curve experiment takes its first steps. At each point of the curve, all dative choices attested so far constitute the training set, while all new dative choices attested in the single next day on which datives are observed constitute the test set. Hence, the first training set is the first day on which the child generated one or more dative constructions; the first test set is derived from the next day the child produced datives. In the second step, the test set of the first step is added to the training data, and the next test set consists of all datives produced by the child on a next day.

2. This may be different for other verbs than *give* or *show*.

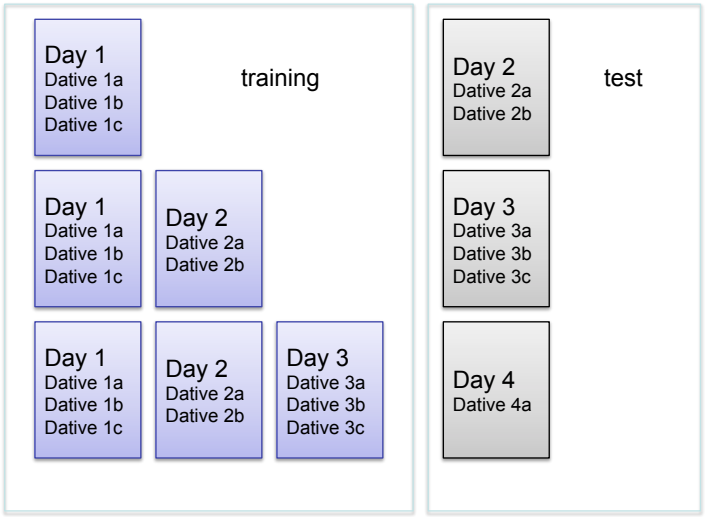

Figure 1. Visualisation of the first steps of a learning curve experiment. In the first step, the training material contains all dative attestations observed in the first day of attestations, and the test material contains all dative attestations found in the next day with datives. In the second, step, the latter material is added to the training set, and the third day of attestations is now the test set.

At each step the incrementally learning memory-based classifier adds the new examples to memory, after which it classifies the new test set, which may only contain one or a handful of attestations. All single predictions per day are recorded as a sequence of predictions and whether these predictions were correct or incorrect. At each point of the curve a correctness score can be produced that aggregates over all predictions so far. At the end of the curve we achieve an aggregate score over all predictions.

The desired outcome of a learning curve experiment is obviously a metric expressing the success of predicting the right choices. In order for individual experimental outcomes to be comparable, they should not be based on different skews in the distribution between the two dative choices. Accuracy (the percentage of correct predictions) will not do, as it is biased to the majority class. When a child would choose one dative construction in 90% of the cases, a classifier trained on that child would easily score 90% accurate predictions by only guessing the majority outcome, while a classifier that is able to attain 80% correct predictions for a child that chooses between the two alternations in a 50%–50% distribution is intrinsically more successful and interesting.

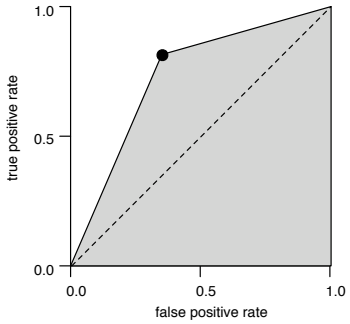

Figure 2. Illustration of the area under the curve (AUC) in the true positive rate–false positive rate space of the outcome of a point classifier (large dot).

To eliminate the effect that class skew may have on our evaluation metric we evaluate our classifier predictions in the learning curve experiments with the area under the curve (AUC) metric (Fawcett 2004). The AUC metric computes, per class, the surface under a curve or a point classifier in the two-dimensional receiver operation characteristic (ROC) space, where the one dimension is the true positive rate (or recall) of predicting the class, and the other dimension is the false positive rate of mispredicting the class. Figure 2 displays the AUC score of the outcome of a classifier (a point classifier as it produces a single score rather than a curve) on a class, depicted by the large dot; the AUC score is the area of the gray surface.

We compute the AUC score of both dative choices, and take the micro-average of the two AUC scores; i.e. each score is weighted by the relative proportion of occurrence of its choice. The resulting number is a score between 0.5 and 1.0 that is insensitive to the skew between the two dative choices in a particular child's data, where 0.5 means baseline performance (random or majority guessing), and 1.0 means perfect prediction.

4 Results

As an illustration of the measurements taken during learning curve experiments, Figure 3 displays the curves for Adam and Nina, the children with most observations. Starting at 100% AUC score, the curves of both children initially drop considerably, and then rise to a score that appears to stabilize, at least for Adam for whom data is available into his fifth year. Later points in the curve are based on more training data.

At the end of each curve, the aggregated score can be measured, which in the best case would be a good approximation of the stabilized score we saw with Adam. Ta-

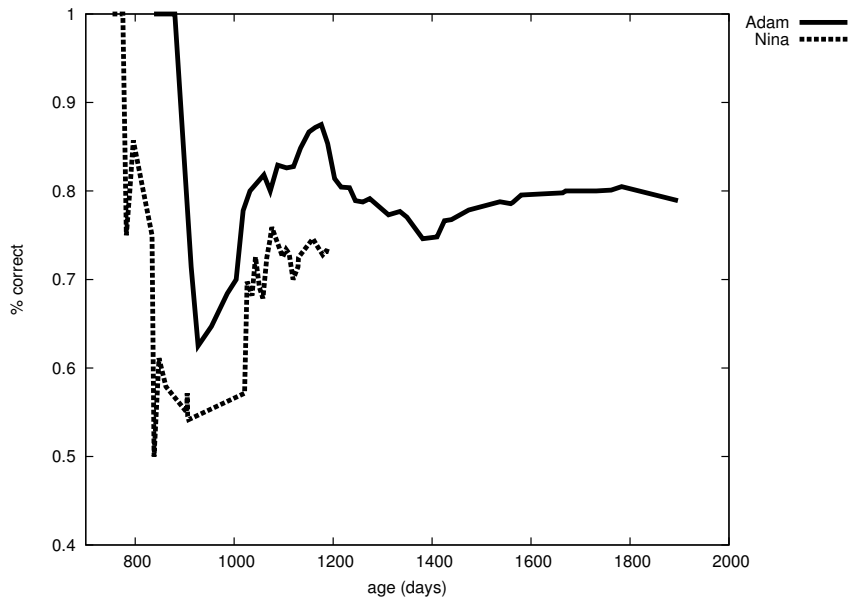

Figure 3. Individual learning curves for Adam and Nina, in terms of AUC scores on predicted dative alternation choices, trained on their own earlier data.

ble 3 lists the aggregated score at the end of the curve for all seven children. Adam's dative choices can be predicted at an AUC score of 0.80, while Nina's choices are predicted with an AUC score of 0.71. For all other children the available data is insufficient to arrive at any above-chance performance.

To arrive at a sufficient amount of data per child we can add the data from all other children to all points of the learning curve, mixing the child's own data with substantially more data from other children. The fourth column of Table 3 shows that this leads to above-chance performance of 0.7 or higher for all children except for Naomi (0.52). However, Adam's score is slightly lower after this mix (0.77 versus 0.80 on Adam's own data).

As De Marneffe *et al.*'s study suggests, it makes sense to predict the children's dative choices from child-directed speech, which represents one of the major sources of language input a child receives. To avoid any effects of alignment (such as the child repeating the caregiver), we constructed training sets for all children that exclude the utterances of their own caretakers. The fifth column of Table 3 lists the AUC scores obtained with this experiment. This leads to improved scores for all children, except for Adam; the score of 0.80 based on his own data is not surpassed.

Table 3. Aggregated AUC scores of MBL at the end of the learning curves of the seven children, training on four different selections of material. Best performances are printed in bold.

		Training on			
Child	# Datives (CDS)	Child only	+ Other children	CDS other children	All
Abe	74	*0.50*	0.84	**0.87**	0.86
Adam	221 (207)	**0.80**	0.77	**0.80**	**0.80**
Naomi	21	*0.50*	0.52	**0.81**	0.58
Nina	146 (443)	0.71	0.74	0.76	**0.79**
Sarah	19	*0.50*	0.83	**0.88**	0.83
Shem	15 (138)	*0.50*	0.74	**0.88**	0.74
Trevor	33	*0.50*	0.72	**0.86**	0.73

Finally, the sixth column of Table 3 displays the scores at the end of the learning curve when all available data is used as additional data during all points of the curve, including all child-directed speech from other children and all other children's data. Surprisingly the advantage of having the maximal amount of training data is not visible in the scores, which are mostly lower, except for Adam (stable at 0.80) and Nina, the other child for which sufficient data was available (0.79).

Overall, the individual scores for all children range between 0.79 and 0.88, which could be considered accurate. For comparison, De Marneffe *et al* report a C score of 0.89 by their aggregate model. The C score (Harrell 2001) is typically used for measuring the fit of regression models, and is to regression what AUC is to classification. It should be noted, though, that the C score is a fit, i.e. a test on the training data, whereas we test un unseen data only.[3] If memory-based learning is applied to classify its training data, its score is trivally 100%, as it memorizes all training exemplars.[4]

It is also possible to reverse the roles in the training and testing regimen, and test the predictive value of children's datives on caregiver datives. This experiment would show how well an child's speech approximates that of adults. Figure 4 displays learning curves (AUC scores) when training on increasing amounts of datives produced by Adam and Nina, tested on the caregiver speech of other children. The score starts out low, then increases, peaks (with both children) and then slowly decreases (with Adam).

3. After reporting on the C score, De Marneffe et al. (2012) note that they do not know whether their model overfits. They then introduce a new experiment on two new children and datives with three verbs: *give*, *show*, and a new verb *bring*, and split the data into a 90% training set and 10% test set. On all three verbs they report a classification accuracy (not AUC, unfortunately) on the test set of 91.2% against a majority baseline of 68.4%. On the new verb *bring* the accuracy is 72.9%.

4. Classification accuracy when testing on the training set may be lower than 100% when identical

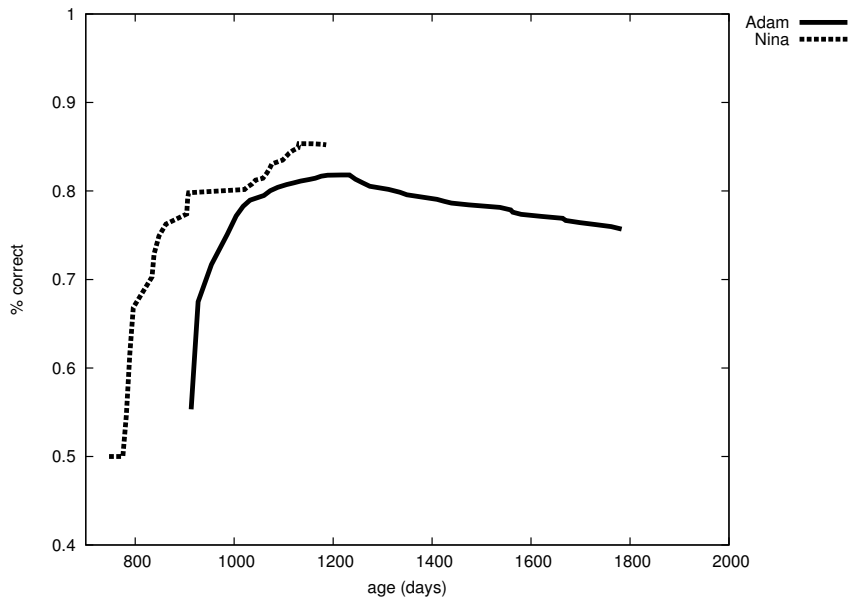

Figure 4. AUC scores on predicting dative alternation choices in child-directed speech from other children, based on increasing amounts of data from Adam and Nina.

To put the outcomes of these two learning curves in perspective, Table 4 compares their aggregrate score against two experiments in which the child-directed speech of Adam and Nina was used, respectively, as training data, and their predictive power was tested on the same test set of other children's child-directed speech. For Adam we observe a higher score, while the number of datives in his child-directed speech is slightly lower (207) than his own datives (221). With Nina, we see that the scores are virtually the same, despite the fact that considerably more caregiver utterances were available (443) than datives produced by Nina herself (146).

5 Discussion

In this contribution we explored the notion of building a predictive computational, exemplar-based model for individual children. Despite the fact that we were only

training exemplars exist with different dative choice labels.

Table 4. Comparison of AUC scores when testing on CDS data from other children, trained either on the child's datives or on the child's caregiver's datives.

Child	# Datives (CDS)	Train child, test cds	Train and test CDS
Adam	221 (207)	0.76	0.84
Nina	146 (443)	0.85	0.86

able to work with a limited number of children for which sufficient data was available, we believe we have delivered a proof of concept: we can model individual learning curves, and when sufficient data is available, the results indicate that models trained on this data have competing generalization performance to aggregate models trained on data from multiple individuals.

What is more, our results indicate that training on other children does not produce the best predictive models. Training on child-directed speech, however, does lead to the overall best generalization performances. This partially confirms De Marneffe *et al.*'s conclusions. Although we used the same data, we cannot directly compare to this work because, as noted before, De Marneffe *et al.* fit their models on the training data, whereas we test on unseen data not included in training.

Finally, we estimated to what extent the data from the children for which we had sufficient data, Adam and Nina, could be used as training data to predict caregiver datives. The comparisons produce slightly different results. Comparing Tables 3 and 4, we observe that Nina's dative choices are harder to predict than Adam's, but they approximate adult caregiver dative choices better. A comparative study of Nina's and Adam's productions may explain this difference, but goes beyond the scope of this paper. We restrict ourselves to noting that we observe more varied predictors in Nina's output than in Adam's, that she uses significantly more pronouns, and that the variance in the length of the themes used by Nina is significantly greater than Adam's.

Overall, both Adam's and Nina's datives can be said to approximate and predict caregiver datives about as accurately as adult data does.

6 Conclusion

Our case study shows that modelling at the level of the individual is possible. Memory-based learning is a suitable method for this type of micro-modelling. It can work with very small amounts of training data, and it can learn incrementally; most non-local regression methods and supervised machine learning methods require complete re-training when training data changes (e.g. when new examples come in). Furthermore,

as an implementation of exemplar-based reasoning it offers a computational, objectively testable, reproducible, and arguably cognitively plausible (Van den Bosch and Daelemans 2013) exemplar-based account of language acquisition and processing.

This proof-of-concept case study suggests several strands of future work. First, different syntactic alternations could be studied in the same way based on the same data, such as the genitive alternation in English. Second, our present study copied the features of De Marneffe et al. (2012), but as remarked before there is some evidence from studies on adult data that the dative alternation can also be predicted with memory-based learning on lexical surface features (words) only (Theijssen 2012). It would be interesting to repeat this study only with the *Theme* and *Recipient* surface lexical features.

As a more general goal, we hope to arrive at a new framework for modeling language production processes in which we can address existing research questions at the individual level, so that we can start to address the contrast between idiolectal data and aggregated data—an issue that has so far been largely theoretical and has been rarely addressed empirically (Louwerse 2004; Mollin 2009).

Acknowledgements

The authors would like to thank Stef Grondelaers and Roeland van Hout for their support, and the participants of the NWASV Workshop in November 2012 for fruitful discussions and feedback. This material is based in part upon work supported by the National Science Foundation under Grant No. BCS-1025602.

References

Aha, D. W., D. Kibler, and M. Albert (1991). Instance-based learning algorithms. *Machine Learning* 6:37–66.

Atkeson, C., A. Moore, and S. Schaal (1997). Locally weighted learning. *Artificial Intelligence Review* 11(1–5):11–73.

Bresnan, J., A. Cueni, T. Nikitina, and R. H. Baayen (2007). Predicting the dative alternation. In G. Bouma, I. Krämer, and J. Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94, Amsterdam, The Netherlands: Royal Netherlands Academy of Arts and Sciences.

Burnard, Lou (2000). Reference guide for the british national corpus (world edition). Technical report, Oxford University.

Campbell, A. and M. Tomasello (2001). The acquisition of english dative constructions. *Applied Psycholinguistics* 22(2):253–267.

Daelemans, W. and A. Van den Bosch (2005). *Memory-based language processing*. Cambridge, UK: Cambridge University Press.

Daelemans, W., J. Zavrel, K. Van der Sloot, and A. Van den Bosch (2010). TiMBL: Tilburg memory based learner, version 6.3, reference guide. Technical Report ILK 10-01, ILK Research Group, Tilburg University.

De Marneffe, M.-C., S. Grimm, I. Arnon, S. Kirby, and J. Bresnan (2012). A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes* 27(1):25–61.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, Hewlett Packard Labs.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Louwerse, M. M. (2004). Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. *Computers and the Humanities* 38(2):207–221.

MacWhinney, B. (2000). *The database*, volume 2 of *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum.

Mollin, S. (2009). "i entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. *Journal of Corpus Linguistics* 14 (3):367–392.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Theijssen, D. (2012). *Making choices: Modelling the English dative alternation*. Ph.D. thesis, Radboud University Nijmegen.

Van den Bosch, A. and W. Daelemans (2013). Implicit schemata and categories in memory-based language processing. *Language and Speech* 56(3):308–326.