Agreement Groups Analysis of Mother-child Discourse

Copyright © 2014
Selected Papers from the 4th UK Cognitive Linguistics Conference
http://uk-cla.org.uk/proceedings
Pages: 52 – 67

LÁSZLÓ DRIENKÓ

dri@t-online.hu

Abstract

We propose a distributional framework for analysing linguistic corpora. The analysis is based on groups of minimally contrasting utterances. Such groups can be considered as representing agreement relations. Agreement groups can be related to the notion of 'frame' used in its various senses in the research literature: item-based phrases (Cameron-Faulkner et al. 2003, Stoll et al. 2009), frequent frames (Mintz 2003, Chemla et al. 2009, Wang and Mintz 2010), flexible frames (St. Clair et al. 2010). Since agreement groups provide a means of representing novel sentences on the basis of sentences already encountered, we tested to what extent they can account for novel utterances in a database. We used the Anne files from the Manchester corpus (Theakston et al. 2001) of the CHILDES database (MacWhinney 2000). It was examined to what extent the agreement groups at a given stage of development can account for the utterances of the immediately following 30-minute session. Agreement groups were extracted from the body of utterances encountered up to the test stage. Examining the data of approximately one year we found that at each developmental stage some 19% - 41% of the utterances of the new session were compatible with the agreement groups extracted from the previous sessions. This amounts to a 6% - 10.3% proportion of novel utterances having been compatible with some groups. The results were slightly improved when a "guessing" mechanism was added. Qualitatively, we also found that the formation of groups may support categorisation, and the actual emergence of grammatical agreement.

Keywords: agreement, categorisation, group formation, distributional analysis, language acquisition

1. Introduction: Agreement groups as distributional analysis

Let us begin with some intuitive remarks in order to see how grammatical agreement and distributional properties may be interrelated in languages. If we replace one word in a grammatical utterance with a word of the same "lexical category", and if the new utterance is grammatical, then we may assume that agreement relations must have been preserved since the new word fulfils the agreement requirements of the original sentence. Furthermore, for unambiguous cases we expect the preservation of the original agreement values. For instance, in example (1) below we replace the noun ‘Adam’ with the nouns ‘Eve’ and ‘people’ – cf. the boldface elements – and get a grammatical and an ungrammatical sentence, respectively. In the ‘Eve’ case the resultant sentence is grammatical and the agreement feature values, 3rd person and singular, are preserved. In the ‘people’ case the resultant sentence is ungrammatical so we do not expect agreement relations to have been preserved. Actually, it is the mismatch of number values – singular versus plural – that causes ungrammaticality.[i]

(1)

Adam hates football → **Eve** hates football (3rd person singular agreement between subject and verb preserved)

→ ***People** hates football (agreement relation spoilt)

When more words are involved in the change the preservation of agreement values cannot be guaranteed. The replacement of ‘Adam’ and ‘hates’ in ‘Adam hates football’, for example, does not affect agreement features for ‘Eve likes football’, however in ‘People hate football’ the values are different.[ii] Cf. (2).

(2)

Adam hates football → **Eve likes** football (agreement values preserved: 3rd person singular)

→ **People hate** football (agreement relation between subject and verb is preserved, but the 3rd-singular is changed to 3rd-plural)

In the case of words with ambiguous lexical categories agreement relations will be preserved when an appropriate lexical item is substituted, i.e. a lexical item whose agreement features are compatible with the requirements of the context. In the Hungarian examples of (3) below, the lexically ambiguous word *vár* ‘castle/waits’ can be replaced with the verb *létezik* ‘exists’ without changing the subject-verb person and number agreement relation. Similarly, when *vár* is replaced with an appropriate noun, e.g. *szentély* ‘temple’, the obligatory number agreement between the subject and the predicate noun is maintained with the same values, i.e. singular for both.

(3)

Az iskola **vár** → Az iskola **létezik**

(the school waits/is waiting (the school exists/is existing

‘the school is waiting(for you)’)		‘the school exists’)
Az iskola **vár**	→	Az iskola **szentély**
(the school castle		(the school temple
‘the school is a castle’)		‘the school is a temple’)

In the case of words with ambiguous agreement values the actual values might change, but the agreement relation is still preserved. In (4), for instance, the noun ‘sheep’ can be considered both plural and singular. In a singular context it can be replaced with a singular noun, e.g. ‘dog’, in a plural context with a plural noun, e.g. ‘dogs’ – agreement values trivially will not change. When ‘sheep’ contributes to the context, however, a word change can alter the actual agreement value, as does the substitution of ‘are’ for ‘is’. Nevertheless, the agreement relation between ‘sheep’ and ‘are’ is still correct.

(4)

The **sheep** is there	→	The **dog** is there (3rd person singular agreement between subject and ‘be’ preserved)
The **sheep** are there	→	The **dogs** are there (3rd person plural agreement between subject and ‘be’ preserved)
The sheep **is** there	→	The sheep **are** there (agreement relation between subject and verb is preserved, but the 3rd-singular value is changed to 3rd-plural)

The above examples were meant to suggest that the agreement properties of “minimal pairs” of grammatical utterances, obtained by changing one word in an utterance, are fairly similar. Next we note that for a given utterance several such minimal pairs can be formed. Consider the sentences in (5).

(5)

Adam hates football

Adam hates **basketball**

Eve hates football

Adam **dislikes** football

Charles hates football

Here each sentence is obtained by altering exactly one word in the original, underlined, sentence ‘Adam hates football’. In each sentence there is an agreement relation between the subject and the verb, invariably with the values 3rd person, singular. Since such a group of sentences represents similar agreement properties[iii], we term it an ‘agreement group’.

Now let us arrange the words of the sentences of our agreement group in a table like below, where each word occurs only once.

Adam	hates	football
Eve	dislikes	basketball
Charles		

Table 1. Agreement group in tabular form.

Evidently, each sentence in the group can be composed by properly concatenating words from different columns of the table. We expect the words in the same column to have similar agreement properties since they belong to the same lexical category and yield grammatical sentences when substituted into the original utterance ‘Adam hates football’. Accordingly, we can say that words within a column belong to the same “agreement category”, and we expect any new combination of words, as dictated by their agreement categories (columns in the table), to be grammatical.

Indeed, we observe that the seven novel sentences in (6) are also compatible with the tabular representation of the agreement group, and that agreement features do not change.[iv]

(6)

Eve hates basketball
Eve dislikes football
Eve dislikes basketball
Charles hates basketball
Charles dislikes football
Charles dislikes basketball
Adam dislikes basketball

Table 1 was created on the basis of five sentences, which means that an originally 5-member group represents 5+7=12 sentences altogether. As sentences grow longer – more columns in the table –, and vocabulary becomes larger – more words in the columns (i.e. more rows) –, the ”generative power” of agreement groups becomes more impressive. For instance, 401 sentences may suffice for the creation of a table with 4 columns and 101 rows, representing $101^4 \approx 10^8$ sentences altogether. Thus, besides carrying immediate information about the grammatical build-up of a language, agreement groups can provide a basis for processing novel sequences. This latter property renders agreement groups a useful tool for modelling processes characteristically involving change over time, notably learning processes, where a training phase typically precedes a given state of competence.

An agreement-groups analysis of a given corpus of linguistic sequences may proceed along the following basic lines:

Extract all possible agreement groups from a body of training sequences.

Assign to every word in the training corpus at least one agreement category, where agreement categories are determined by the columns of (hypothetical) tabular representations of agreement groups.

Calculate the proportion of novel sequences in the test corpus that are 'compatible' with the agreement groups gained from the training corpus.

Extracting all possible agreement groups may naturally result in redundancy since there may be groups with identical members. Such groups can be eliminated during computation, however we believe that redundancy should be a component of a realistic model. Corpus a) in (7), for an example, yields agreement groups b) - f), of which e) and f) contain the same sentences, and b) consists of the sentences of c) and d).

(7)

a)

Adam hates football
Adam hates basketball
Eve hates football
Boys hate football
Girls hate football

b)

Adam hates football
Adam hates basketball
Eve hates football

c)

Adam hates basketball
Adam hates football

d)

Eve hates football
Adam hates football

e)

Boys hate football
Girls hate football

f)

Girls hate football
Boys hate football

Note that the mere inspection of agreement groups may reveal some regularities in the data. The first word in every sentence of (7) represents a human being, the first words in b) - d) are proper nouns, the last words refer to sport, b) - d) involve 3rd-singular agreement whereas e) - f) involve 3rd -plural, etc.

Computationally, the compatibility of novel sentences with agreement groups is determined by a mapping process. A novel sentence can be mapped onto an agreement group if the agreement categories of the individual words license a category sequence representing an agreement group. Let us assume the following category assignment for the words in (7) a), our example corpus (letters identify groups, numbers refer to positions (table columns)):

(8)

Adam{b1, c1, d1}	Eve{**b1**, d1}	hates{**b2**, c2, d2}
football{b3, c3, d3, e3, f3}	basketball{**b3**, c3}	hate{e2, f2}

As the boldface items indicate, novel sequence ‘Eve hates basketball’ can be mapped on group (7) b) because it licenses category sequence *b1b2b3* which in turn symbolises group b).

The mapping algorithm can be extended in such a way that allows for some kind of “guessing” as to the categorical status of unknown words, which broadens the scope of novel sequence processing. For instance, ‘People hate football’ can be mapped on group e) or f) if ‘people’ is supposed to have category e1 or f1 (or both).

The “precision” of our method is determined by various factors. Ideally, the training corpus should be “perfectly grammatical”, words should be unambiguous, and the “same lexical category” requirement should be fulfilled in forming minimal pairs for agreement groups. Real life corpora are seldom ideal, yet agreement groups might still be useful. First of all because we expect violations of the idealness criteria in only a fraction of the training set. Secondly, because the main effect of incorrect grouping – the consequence of non-idealness – is that agreement groups will “overgeneralise”, i.e. they will be compatible with some ungrammatical novel sentences in the test set, if there are any. However, they remain compatible with the same grammatical sequences. A third point is that the training set can be corrected “manually”: ungrammaticality can be removed, ambiguity can be resolved by assigning different forms to different meanings (e.g. *vár_1* versus *vár_2*), or lexical categories can be directly indicated (e.g. *vár_verb*, *vár_noun*). Furthermore, for practical applications, agreement groups can also be corrected either manually or by adding new components to the grouping algorithm.

We remark here that it is also logical to suppose some error-correcting mechanisms to affect natural language acquisition. Consider the examples in (9) from Pinker (1979:240).

(9)

Hottentots must survive

Hottentots must fish

Hottentots eat fish

Hottentots eat rabbits

The possible agreement groups are the following:

(10)

a)

Hottentots must survive

Hottentots must fish

b)

Hottentots must fish

Hottentots must survive

Hottentots eat fish

c)

Hottentots eat fish

Hottentots must fish

Hottentots eat rabbits

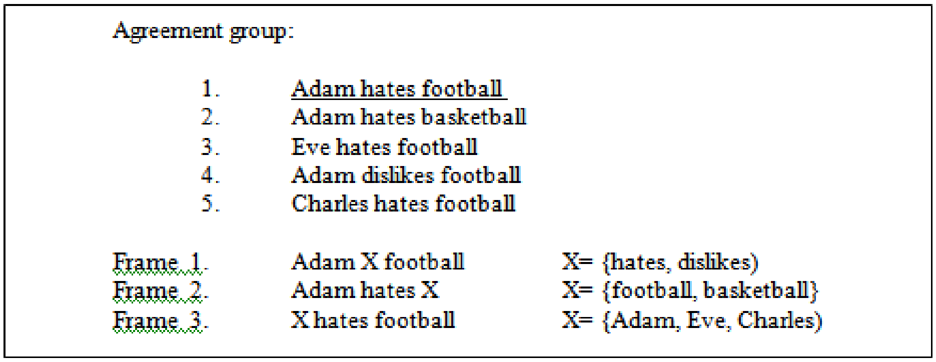
d)

Hottentots eat rabbits

Hottentots eat fish

Due to the ambiguity of ‘fish’, groups b) and c) are erroneous as they would license the incorrect ‘Hottentots eat survive’ and ‘Hottentots must rabbits’, respectively. A group-correcting mechanism could improve the situation by deleting ‘Hottentots eat fish’ from b) and ‘Hottentots must fish‘ from c). Alternatively, groups b) and c) could be deleted completely from the learner’s memory.

Agreement groups constitute a kind of distributional approach insofar as the grouping of utterances is determined by the distribution of words they consist of. Distributional methods in linguistics date back at least to Harris (1951). For Harris the distribution of a linguistic item was determined by all the contexts, or “environments” for that particular item. Kiss (1973) proposed a word categorisation model based on cluster analysis which was extended for larger corpora and computational resources by Redington et al. (1998). Finch et al. (1995) adopted a similar method to assign categories to word sequences, i.e. to phrases. Such clustering methods typically operate with “context vectors” as determined by the neighbouring elements of a target item. Mintz (2003) used a different, more direct, formalisation of context. In his work, the immediately preceding and succeeding words provide the context or “frame” for categorising. Mintz employs “frequent frames”, i.e. contexts with a frequency larger than an arbitrarily defined threshold. Weisleder and Waxman (2010) consider, besides Mintz’s “mid-frames”, the usefulness of “end-frames”, where the utterance-end marker constitutes an informative element. St. Clair et al. (2010) claim that “flexible frames” exploiting bigram information within frequent frames are more optimal for categorisation than just frequent frames. A kind of framing effect in language acquisition was reported by Cameron-Faulkner et al. (2003) who pointed out that mothers speaking to their children use a rather limited set of item-based phrases, these phrases being framed by their initial words. Such findings were confirmed cross-linguistically by Stoll et al. (2009). Below we sketch how agreement groups are related to frames. Cf. Figure 1.



Agreement group:

1. Adam hates football
2. Adam hates basketball
3. Eve hates football
4. Adam dislikes football
5. Charles hates football

Frame 1.	Adam X football	X= {hates, dislikes)
Frame 2.	Adam hates X	X= {football, basketball}
Frame 3.	X hates football	X= {Adam, Eve, Charles)

Figure 1. Agreement group as superposition of frames.

By substituting words of category X into the appropriate position (“slot”) in a frame we get a subset of the agreement group. Frame 1, ‘Adam X football’, yields sentences 1 and 4, i.e. ‘Adam hates football’ and ‘Adam dislikes football’. Frame 2 licenses sentences 1 and 2, whilst Frame 3 symbolises sentences 1, 3, 5. By forming the union of the utterances licensed by the individual frames we have sentences 1, 2, 3, 4, 5, i.e. the whole group. Thus the agreement group represents a “superposition” of frames.

Wang and Mintz (2010:p. 6) propose that “grammatical relations between words are more consistent in individual frequent frames than in bigrams” and that “words within a frequent frame are especially "close" syntactically” (p. 8). This is in accordance with our view that agreement groups represent syntactic (namely, agreement) relations. Bannard and Matthews (2008) suggest that children tend to store word sequences in memory during language acquisition. It could be hypothesized that such word sequences can form the basis of sentence patterns, and that the appropriate grouping of the stored sequences might be a principal element in the emergence of linguistic behaviour. Thus the agreement group idea as discussed in the present work might also be viewed as a kind of model of the organizational processes concerning stored sequences. Agreement groups can also be regarded as symbolising linguistic patterns representing agreement relations in the generalised sense of Drienkó (2004a, b; 2009).[v] It can also be shown that agreement groups, understood as linguistic patterns, are computationally learnable.[vi]

2. The experiment

The data of one particular child, Anne, were chosen from the Manchester corpus (Theakston et al. 2001) of the CHILDES database (MacWhinney 2000). The age span was 1;10.7 – 2;9.10. The files were converted to simple text format, annotations were removed, together with punctuation symbols. Mother and child utterances were not separated, the data were considered as representing a single "mother-child language". Noise was not removed, as it is part of the learning process, that is utterances containing 'xxx' were allowed. Similarly, ungrammaticality cannot be separated from learning, so ungrammatical utterances were also included. The mother-child language at a certain point in time was seen as a language compatible with the data that far, and was not expected to fulfil requirements of grammaticality coming from the language expert's knowledge. "Too long" utterances were removed since long sequences tend to form one-member "groups" which play no part in the processing of novel utterances. The upper bound for utterance length was set to five words.[vii] One-word utterances were also excluded as meaningless from the analysis point of view. Each 60-minute mother-child session was considered as a point in time of linguistic development, alternatively, as a developmental stage. Each point represented the linguistic knowledge acquired up to that point. It was tested to what extent the agreement groups at a given stage can account for the utterances of the immediately following 30-minute session. Agreement groups were extracted from the body of all the utterances, meeting the above criteria, encountered up to the test stage. Each utterance had its own group. It was investigated to what extent the utterances of the next session could be "mapped" on the already existing agreement groups.

3. Results

The experiment revealed the following facts:

1. At every developmental stage there are novel utterances compatible with some agreement group. Additionally, extended mapping (guessing of categories) may improve processing.

2. Agreement categories (as modelled by our hypothetical table columns) may relate quite naturally to notions like "lexical category", and "semantic category".

3. Traces of grammatical agreement can be found in agreement groups.

Below we go through these points in more detail.

Our results concerning the proportion of novel utterances that were compatible with at least one agreement group are visualised for each developmental stage in Figure 2. The range of mapped novel utterances is 6% - 10.3%. It is 6% - 8.9% for the first part of the diagram (for the first 14 sessions) and 7.3% - 10.3% for the later developmental stages, which suggests a slight increase. The figure also displays that some 19% - 41% of the total amount of test utterances, both novel and non-novel, were compatible with at least one group. For non-novel utterances compatibility actually means that the very same utterance had already been heard by the child, so the utterance has its own group and may be a member in others, which makes mapping trivial. The Appendix gives further details on the experiment conditions.

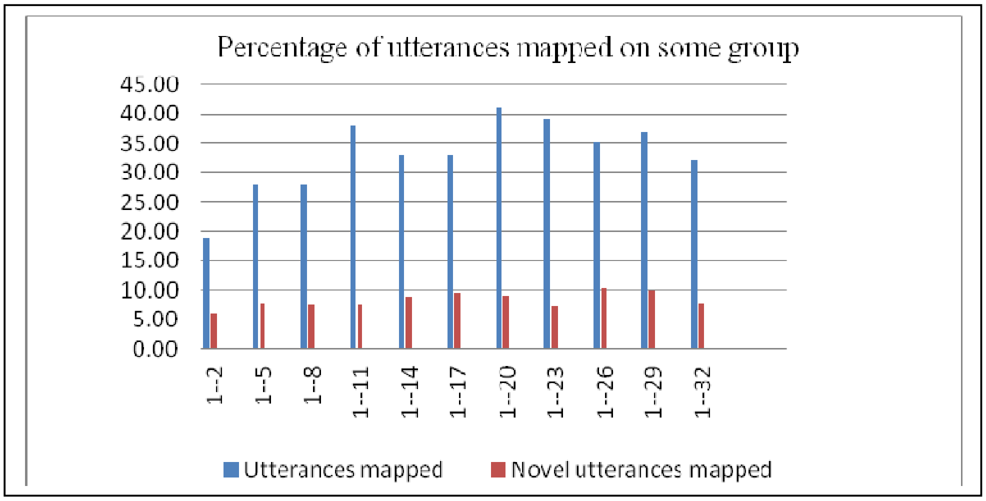
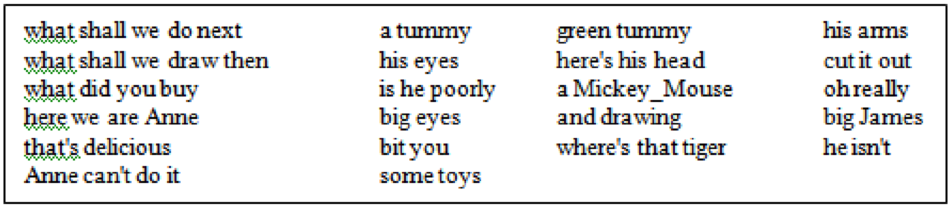


Figure 2. Proportions of utterances mapped on at least one group against developmental stages. Developmental stages are expressed in file numbers. E.g.: 1-17 corresponds to the first seventeen 60-minute sessions.

Table 2 lists the novel utterances that were compatible with the final set of agreement groups at stage 1-32. Table 3 shows the corresponding agreement groups for some of the novel utterances.



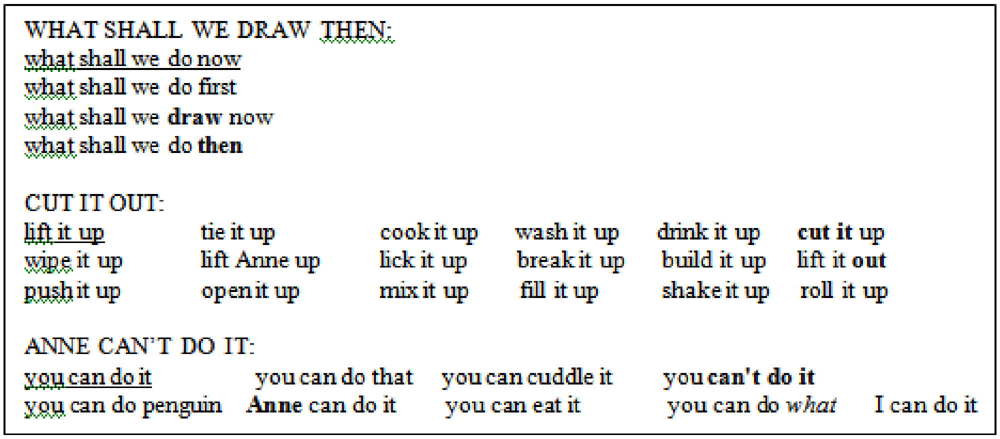
what shall we do next
what shall we draw then
what did you buy
here we are Anne
that's delicious
Anne can't do it

a tummy
his eyes
is he poorly
big eyes
bit you
some toys

green tummy
here's his head
a Mickey_Mouse
and drawing
where's that tiger

his arms
cut it out
oh really
big James
he isn't

Table 2. Novel utterances at stage 1-32.



WHAT SHALL WE DRAW THEN:
what shall we do now
what shall we do first
what shall we **draw** now
what shall we do **then**

CUT IT OUT:

lift it up	tie it up	cook it up	wash it up	drink it up	**cut it** up
wipe it up	lift Anne up	lick it up	break it up	build it up	lift it **out**
push it up	open it up	mix it up	fill it up	shake it up	roll it up

ANNE CAN'T DO IT:

you can do it	you can do that	you can cuddle it	you **can't do it**	
you can do penguin	**Anne** can do it	you can eat it	you can do *what*	I can do it

Table 3. Some novel utterances with the corresponding groups.

We noted earlier that the mapping algorithm can be extended so that some degree of "guessing" as to the categorical status of unknown words can be involved. As a supplement to our main experiment we tested the body of agreement groups gained from the 32 Anne-sessions against the last four files of the Anne-data (files 33a-34b). As Figure 3 shows, the extended mapping mechanism resulted in somewhat better processing. The example values on the diagram reveal how a few percent rise in the number of novel utterances processed (from 8.1% to 12.2%) effects a similar increase in the overall number of processible utterances (from 37% to 42%). Improved processing, in turn, suggests that a capacity for handling unknown items might be a useful component of a realistic model.

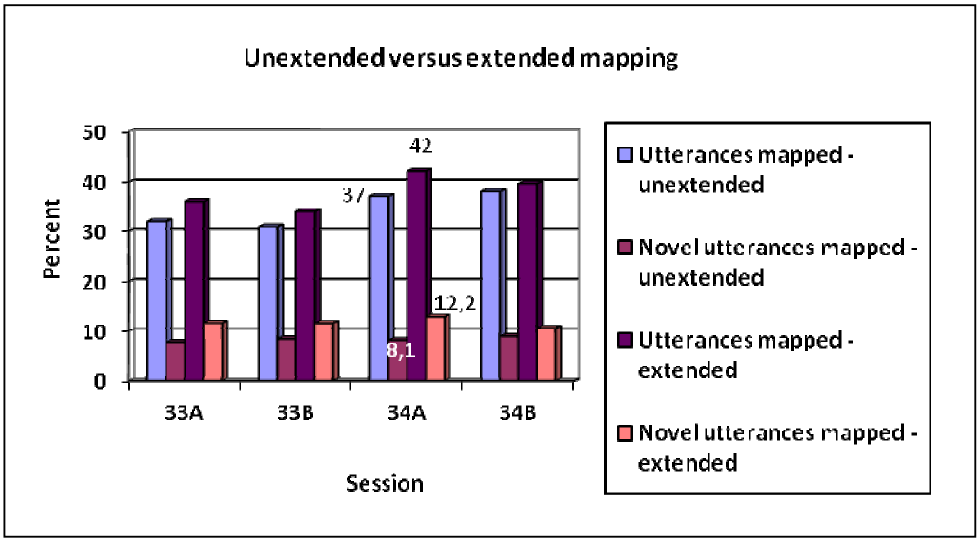
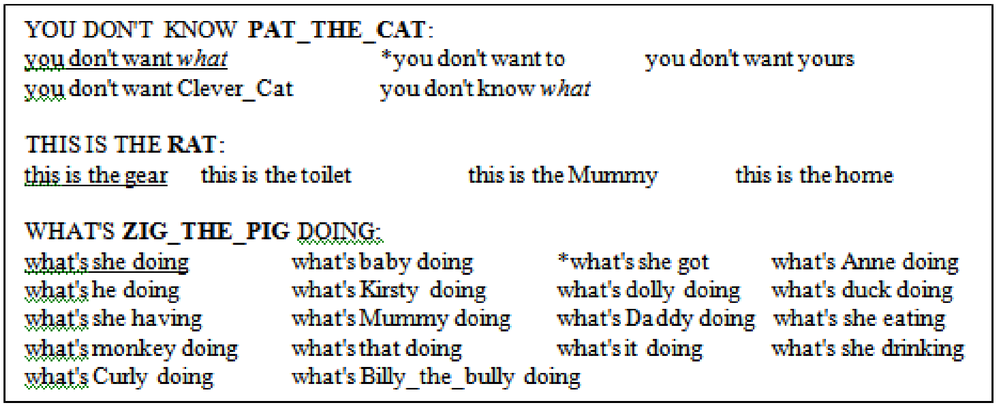


Figure 3. Extension of the mapping mechanism. Extended mapping (category guessing) yields better processing results.

Table 4 provides samples of novel utterances together with the groups they could be mapped on via extended mapping. For an example, in ‘what's Zig_the_Pig doing’ ‘Zig_the_Pig’ is an unknown word but it fits nicely into the ‘what's X doing’ frame which suggests category X for ‘Zig_the_Pig’. Note that sentences ‘what's Zig_the_Pig eating/drinking/having’ would also be mappable. Even ‘what's Zig_the_Pig got’ would be appropriate, where ‘'s’ stands for ‘has’. We also note in passing that such guessing mechanisms remain operative for the competent speaker of the language. On hearing a sentence like ‘what's Zaxvul doing’ the speaker can only guess that ‘Zaxwul’ must be some name, some person or animal, something capable of doing something, etc.



YOU DON'T KNOW **PAT_THE_CAT**:

you don't want *what*	*you don't want to	you don't want yours
you don't want Clever_Cat	you don't know *what*	

THIS IS THE **RAT**:

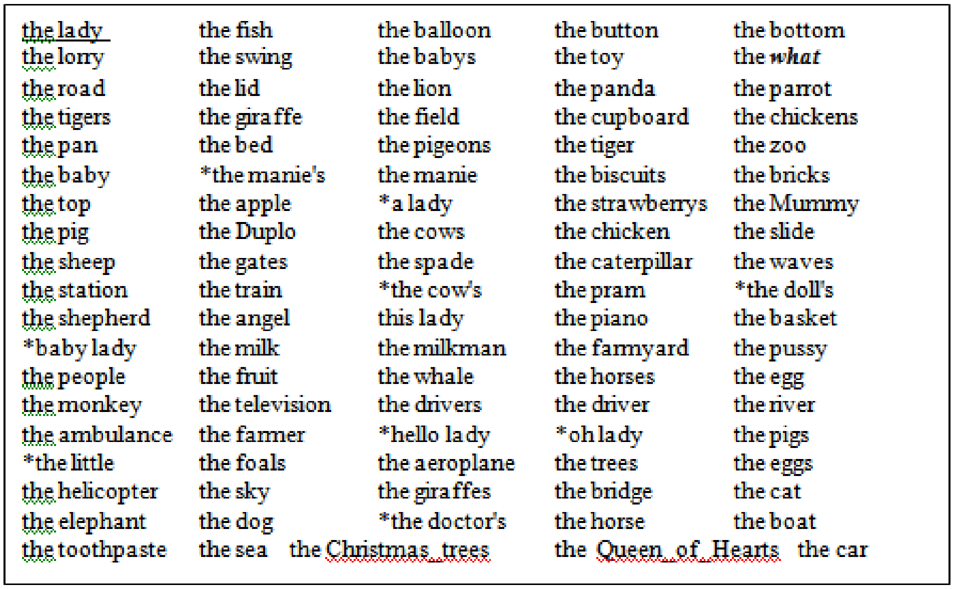
this is the gear	this is the toilet	this is the Mummy	this is the home

WHAT'S **ZIG_THE_PIG** DOING:

what's she doing	what's baby doing	*what's she got	what's Anne doing
what's he doing	what's Kirsty doing	what's dolly doing	what's duck doing
what's she having	what's Mummy doing	what's Daddy doing	what's she eating
what's monkey doing	what's that doing	what's it doing	what's she drinking
what's Curly doing	what's Billy_the_bully doing		

Table 4. Novel utterances with corresponding agreement groups. (Extended mapping.) Unfamiliar words are printed in bold.

Agreement categories (columns of hypothetical tables for agreement groups) may subserve more traditional aspects of categorisation. The ‘this is the gear’ group in Table 4, for instance, represents a ‘this is the X’ frame with X symbolising nouns. The ‘what's she doing’ group quite homogeneously instantiates a ‘what's X Y’ pattern, where X stands for a nominal item whilst Y denotes an ‘-ing’ verb, with ‘what's she got’ as the only exception (as the asterisk indicates). The ‘you don't want what’ group is a bit more heterogeneous. In the lexical sense ‘what’, ‘to’, ‘yours’, and ‘Clever_cat’ seem to belong to separate categories. However, in a broader sense, and with the exception of ‘to’, these words can also be labelled as nominal. Note that the explicit emergence of ‘what’ (or other question word) in a group may be of special interest because it allows direct association of syntactic categories and utterance positions. The italicised ‘what’ in ‘you can do what’ in Table 3 is another possible example of nominal category association. Table 5 shows that the ‘the X’ construction is fairly effective for categorising nouns.[viii]



the lady
the lorry
the road
the tigers
the pan
the baby
the top
the pig
the sheep
the station
the shepherd
*baby lady
the people
the monkey
the ambulance
*the little
the helicopter
the elephant
the toothpaste

the fish
the swing
the lid
the giraffe
the bed
*the manie's
the apple
the Duplo
the gates
the train
the angel
the milk
the fruit
the television
the farmer
the foals
the sky
the dog
the sea

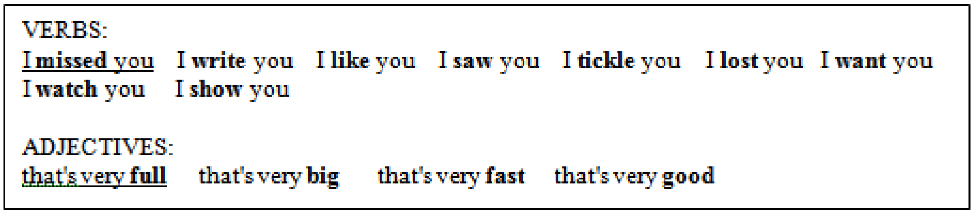
the balloon
the babys
the lion
the field
the pigeons
the manie
*a lady
the cows
the spade
*the cow's
this lady
the milkman
the whale
the drivers
*hello lady
the aeroplane
the giraffes
*the doctor's
the Christmas trees

the button
the toy
the panda
the cupboard
the tiger
the biscuits
the strawberrys
the chicken
the caterpillar
the pram
the piano
the farmyard
the horses
the driver
*oh lady
the trees
the bridge
the horse
the Queen of Hearts

the bottom
the ***what***
the parrot
the chickens
the zoo
the bricks
the Mummy
the slide
the waves
*the doll's
the basket
the pussy
the egg
the river
the pigs
the eggs
the cat
the boat
the car

Table 5. A group representing a 'the X' nominal phrase.

The 'I X you' and 'that's very X' constructions of Figure 4 exemplify frames categorising verbs and adjectives.

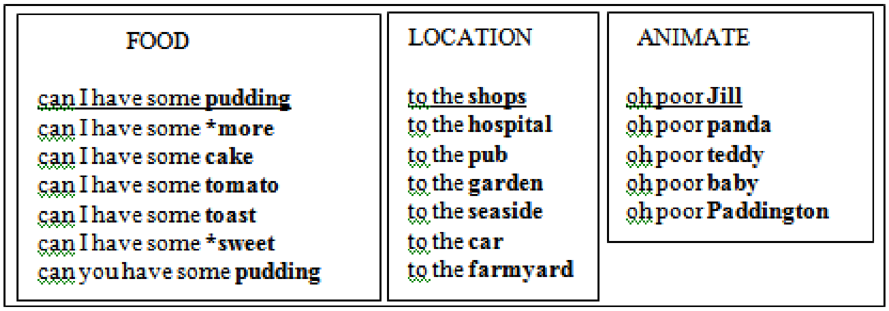


VERBS:
I **missed** you   I **write** you   I **like** you   I **saw** you   I **tickle** you   I **lost** you   I **want** you
I **watch** you   I **show** you

ADJECTIVES:
that's very **full**   that's very **big**   that's very **fast**   that's very **good**

Figure 4. Agreement groups facilitating the categorisation of verbs and adjectives.

Categorisation by agreement groups may have a semantic orientation. For instance, the frames the groups in Figure 5 represent require nouns with specific semantic content: the boldface words of the 'can I have some X' group refer to some food, 'to the X' specifies some location, and 'oh poor X' has humans or animals, i.e. personified animate beings, as referents.

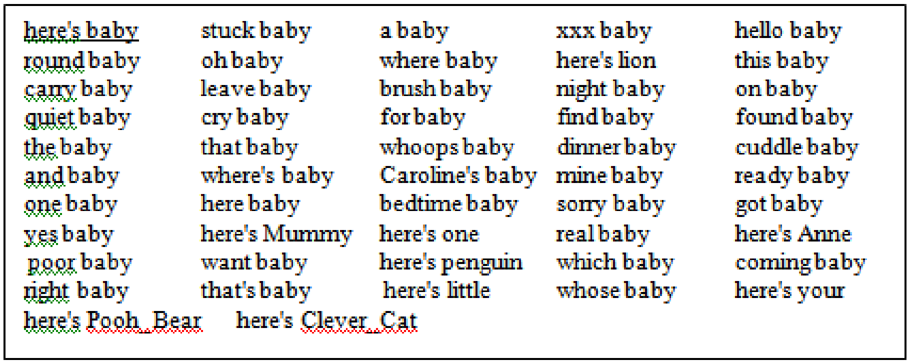


FOOD	LOCATION	ANIMATE
can I have some **pudding**	to the **shops**	oh poor **Jill**
can I have some ***more**	to the **hospital**	oh poor **panda**
can I have some **cake**	to the **pub**	oh poor **teddy**
can I have some **tomato**	to the **garden**	oh poor **baby**
can I have some **toast**	to the **seaside**	oh poor **Paddington**
can I have some ***sweet**	to the **car**	
can you have some **pudding**	to the **farmyard**	

Figure 5. Agreement categories as semantic concepts.

Conceptually, we relate our analysis to the notion of ‘agreement’. Such a relationship can gain further, empirical, justification if traces of grammatical agreement can be detected in the data. Despite the fact that agreement relations are relatively poorly represented in English, at least rudiments of them are readily detectable in the groups emerging in our experiment. Consider, for some examples, the groups in Table 4. Every sentence in the ‘you don't want what’ group is grammatical primarily because ‘you’ and ‘don't’ have the same person-number values, i.e. ‘she don't’ or ‘you doesn't’ would not be correct. By the same token, ‘this are the gear’, or ‘these is the gear’ could spoil grammaticality in a rigorous sense for the ‘this is the gear’ group, just as ‘what's they doing’, or ‘what're she doing’ would mean ungrammatical members for the ‘what's she doing’ group.

In connection with the “precision” of analysing real-life corpora we hinted above that agreement groups might not be “perfect”. Indeed, we found some quite strange-looking groups. However, it could be observed that suboptimal behaviour is most typical of groups of two-word long utterances. This is quite in line with the findings on frames by Mintz (2003), Weisleder and Waxman (2010), and St. Clair et al. (2010) pointing to a possible assumption that the larger portion of context is considered, the higher accuracy of lexical categorisation is achieved: in two-word long utterances there is only bigram information, the context is just one word. The high degree of heterogeneity of the ‘here's baby’ group in Table 6, for example, is largely due to the diversity of syntactic categories. An ‘X Noun’ pattern may yield several utterance types: Verb Noun (e.g. carry/leave/cry/got baby), Adjective Noun (e.g. real/poor baby), Noun Noun (e.g. bedtime/dinner baby), Preposition Noun (e.g. for/on baby), Interjection Noun (e.g. oh/whoops baby), etc. That the ‘here's X’ frame component of the same group is perhaps more restrictive with respect to syntactic categories is also a confirmation of the assumption that larger contexts result in better categorisation since ‘here's’ actually means a two-word context, i.e. ‘here is’, for X.



here's baby
round baby
carry baby
quiet baby
the baby
and baby
one baby
yes baby
poor baby
right baby

stuck baby
oh baby
leave baby
cry baby
that baby
where's baby
here baby
here's Mummy
want baby
that's baby

a baby
where baby
brush baby
for baby
whoops baby
Caroline's baby
bedtime baby
here's one
here's penguin
here's little

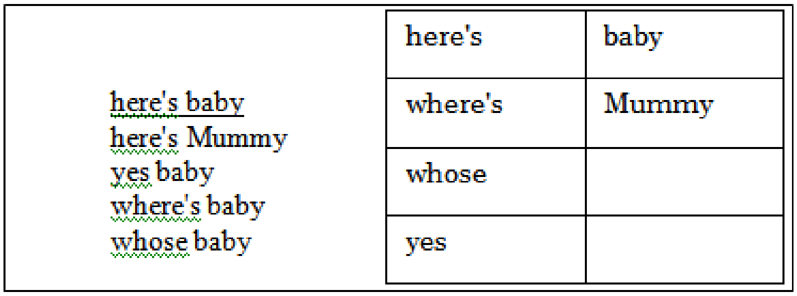
xxx baby
here's lion
night baby
find baby
dinner baby
mine baby
sorry baby
real baby
which baby
whose baby

hello baby
this baby
on baby
found baby
cuddle baby
ready baby
got baby
here's Anne
coming baby
here's your

here's Pooh_Bear here's Clever_Cat

Table 6. Heterogeneous group.

Note that syntactic/lexical heterogeneity does not necessarily spoil the generalisation properties of a group. Inasmuch as group members are grammatical minimal pairs of a specified utterance and no ambiguity is involved, we may "generate" grammatical novel utterances even if the "same lexical category" condition does not hold. To see this consider a fragment of the 'here's baby' group and the corresponding hypothetical table in Figure 6. Clearly, the items in the first column fall into several lexical categories. The consequence is that the novel utterances 'where's Mummy', 'whose Mummy', and 'yes Mummy' represent different utterance types, just like the original members of the group. The agreement relations of any novel utterance correspond to the type the utterance belongs to. In the 'here's X', and 'where's X' cases there is person and number agreement, while 'yes X' and 'whose X' require no such agreement. Thus lexically heterogeneous groups may be characteristic of intermediate developmental stages where syntactic patterns are not yet fully dissociated with respect to their agreement requirements.

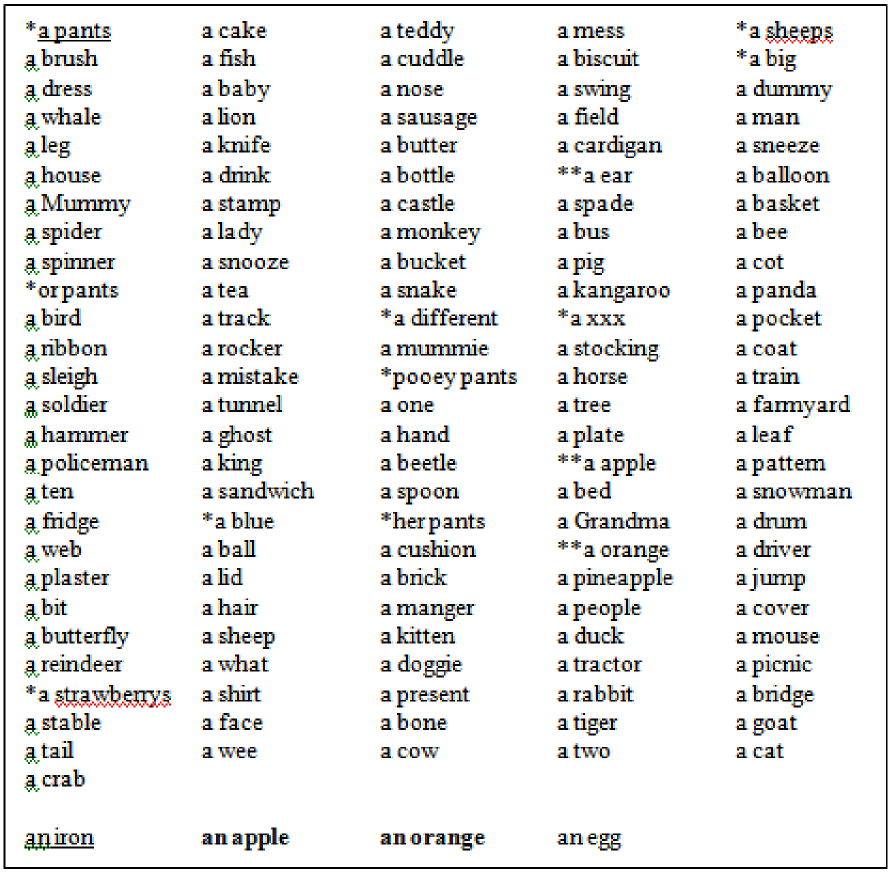


here's baby
here's Mummy
yes baby
where's baby
whose baby

here's	baby
where's	Mummy
whose	
yes	

Figure 6. Agreement group and corresponding table representing various utterance types.

The group in Table 6 licenses 'a Anne' which is "approximately" correct: if understood as a determiner-noun sequence it can be legal since articles may sometimes co-occur with proper nouns. Phonologically, however, 'an Anne' would be desirable, which suggests a phonological dimension for group formation. In fact, we found traces of ‘a’-‘an’ dissociation in the data. Cf. the examples of Table 7.



*a pants	a cake	a teddy	a mess	*a sheeps
a brush	a fish	a cuddle	a biscuit	*a big
a dress	a baby	a nose	a swing	a dummy
a whale	a lion	a sausage	a field	a man
a leg	a knife	a butter	a cardigan	a sneeze
a house	a drink	a bottle	**a ear	a balloon
a Mummy	a stamp	a castle	a spade	a basket
a spider	a lady	a monkey	a bus	a bee
a spinner	a snooze	a bucket	a pig	a cot
*or pants	a tea	a snake	a kangaroo	a panda
a bird	a track	*a different	*a xxx	a pocket
a ribbon	a rocker	a mummie	a stocking	a coat
a sleigh	a mistake	*pooey pants	a horse	a train
a soldier	a tunnel	a one	a tree	a farmyard
a hammer	a ghost	a hand	a plate	a leaf
a policeman	a king	a beetle	**a apple	a pattern
a ten	a sandwich	a spoon	a bed	a snowman
a fridge	*a blue	*her pants	a Grandma	a drum
a web	a ball	a cushion	**a orange	a driver
a plaster	a lid	a brick	a pineapple	a jump
a bit	a hair	a manger	a people	a cover
a butterfly	a sheep	a kitten	a duck	a mouse
a reindeer	a what	a doggie	a tractor	a picnic
*a strawberrys	a shirt	a present	a rabbit	a bridge
a stable	a face	a bone	a tiger	a goat
a tail	a wee	a cow	a two	a cat
a crab				
an iron	**an apple**	**an orange**	an egg	

Table 7. Two groups representing a dissociation of indefinite nominal phrases with respect to the phonological properties of the nouns. Boldface items are the correct ‘an’ versions of phonologically incorrect (marked ‘’) utterances in the ‘a X’ group.**

The need for further refinement of agreement groups is also reflected in the fact that groups may have ungrammatical members. For instance, the grammatical status of ‘here's little’, or ‘mine baby’ in Table 6 is rather dubious. In our framework the improvement of agreement groups may be associated with a simple correction mechanism. First, erroneous/non-fitting group members can be deleted for making the group more correct or consistent. Secondly, if individual deletions cannot better the situation, the whole group can be deleted from memory. As we allow for – we believe – reasonable redundancy, the loss of a single group cannot be fatal. A (or the most) similar group can take its place. A similar and more correct group will license similar and more correct sequences. The evaluation criteria of individual groups may rest on effectiveness: the more ungrammatical utterances the group is compatible with (i.e. is capable of “generating”), the more degree of improvement it needs, the ultimate degree being total deletion.

From a language acquisition point of view, perhaps, “(un)grammatically” here should be understood as e.g. “cognitive/communicative (in)effectiveness” since the acquisition of a first language cannot be guided by explicit grammaticality instructions. To put it simply, then, a group is more incorrect if the utterances it licenses are less intelligible. By “intelligible” we mean the capacity of the learner, and/or the listener to accommodate utterances in their cognitive systems. When an utterance like ‘Hottentots must rabbits’ is unintelligible to the listener, he or she can give explicit feedback to the learner, which, in turn, may be explicit feedback on group effectiveness. Alternatively, the learner can “discover” that a group licensing ‘Hottentots must rabbits’ is incorrect since ‘rabbits’, unlike e.g. ‘survive’ or ‘fish’, is associated with ‘what’ most of the time, therefore it cannot occupy an utterance position assigned to verbs.

4. Conclusions and future work

The present study outlined a distributional framework based on the notion of agreement groups. Our analysis revealed that at each developmental stage some 19% - 41% of the utterances of the new session were compatible with the agreement groups extracted from the previous sessions. The compatible novel utterances were in the 6% - 10.3% range. The proportion of novel utterances increased slightly with time but this needs to be confirmed by future research. We also found some evidence that extended mapping, i.e. category guessing, may improve processing.

Our qualitative inspection of the actual agreement groups led us to two additional conclusions. Firstly, agreement categories, as defined in terms of positions within the utterances of individual groups, may serve as a basis for words to dissociate into “higher-level” categories with respect to their lexical, semantic, or phonetic properties. Secondly, traces of grammatical agreement can be found in agreement groups since group members do exhibit similar agreement properties. Inconsistencies, or heterogeneities, within groups – we propose – may be regarded as transitional assuming a suitable error correction mechanism.

There may be several frontiers for widening the horizons of the kind of analysis we presented here. Other languages can provide cross-linguistic contrast for evaluating the findings reported. In fact, some preliminary experiments with Spanish and Hungarian data seem to indicate that agreement groups may provide relevant information cross-linguistically as well. Another pilot investigation of ours appears to mark out another direction of possible further improvement: if we allow agreement groups to account for parts of larger novel sentences, a larger coverage of novel data can be attained. That is, speech fragments acquired separately may account for novel combinations of such fragments in longer utterances.

We hope that our results may qualify as some additional evidence for the importance of distributional research. Furthermore, due to a theoretical connection between agreement groups and linguistic agreement patterns as introduced in Drienkó (2004a, b; 2009), and to research findings suggesting that children tend to store word sequences in memory during language acquisition (Bannard and Matthews 2008), we also believe that agreement groups might possibly be a useful tool in advancing linguistic theory.

Notes

[i] Of course, if we consider ‘people’ to belong to a different lexical category – common noun as opposed to the proper nouns ‘Adam’ and ‘Eve’– we do not expect agreement feature matching in the first place since the “same lexical category” replacement requirement is not fulfilled. We would not like to go into much detail about the possibility of a precise definition for lexical categories. An intuitive notion of lexical, or grammatical, category will suffice in setting the context for our findings.

[ii] For simplicity, we regard verbs without an ’-s’ suffix as 3rd person plural.

[iii] In the general case, by ’agreement’ we refer to the generalised notion of ’agreement’ proposed in Drienkó (2004a, b; 2009). In that sense, feature correlation of any nature – phonological, morphological, even configurational – can be regarded as an agreement relation.

[iv] Note the importance of our “one-word-difference” requirement for agreement groups. If, for instance, ‘Adam hates football’ and ‘People hate football’ belonged to the same group, the incorrect ’Adam hate football’ and ‘People hates football’ would also be licensed by the corresponding table:

Adam	hates	football
people	hate	

[v] Since Drienkó differentiates between recursive and non-recursive patterns, it is more precise to say that agreement groups symbolise non-recursive agreement relations. ‘Recursive’ means that certain part(s) of the pattern can be repeated. E.g. ‘The boy likes, the girls hate football’ can be mapped on a recursive pattern like *(Det N V)*^{i} *N* where *(Det N V)* can be repeated arbitrarily many, *i*, times.

[vi] Drienkó, L. (2011). Inference of non-recursive agreement patterns: theory, and application to mother-child speech. Unpublished manuscript.

[vii] Note that utterance length limitations for the training set actually reduce the potential of the mapping mechanism. This is alleviated by the same limitations for the test set, where the exclusion of too long utterances enhances the mapping rate.

[viii] Words are spelt as in the CHILDES transcripts.

References

Bannard, C., Matthews, D. (2008). Stored Word Sequences in Language Learning. *Psychological Science,* 19(3), 241-248.

Cameron-Faulkner, Th., Lieven, E., Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*. 27. 843-873.

Drienkó, L. (2004a). Agreement Mapping System Approach to Language. *Journal of Language and Linguistics*. Vol. 3. No. 1. 38-61.

Drienkó, L. (2004b). Outlines of Agreement Syntax. *Journal of Language and Linguistics*. Vol. 3. No. 2. 154-181.

Drienkó, L. (2009). *A linguistic agreement mapping-system*. Unpublished PhD dissertation, ELTE University, Budapest.

Finch, S., Chater, N., Redington, M. (1995). Acquiring syntactic information from distributional statistics. In: Levy, JP, Bairaktaris, D, Bullinaria, JA, Cairns, P, (eds.) *Connectionist models of memory and language*. (229 - 242). UCL Press: London.

Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago, IL, US: University of Chicago Press.

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation*, 7, l-41.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, Volume 90, Issue 1, pp. 91-117. doi:10.1016/S0010-0277(03)00140-9

Pinker, S. (1979). Formal models of language learning *Cognition,* 7, 217-283.

Redington, M., Chater, N., Finch, S. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science* Vol. 22 (4) pp. 425-469.

St. Clair, M. C., Monaghan, P., Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*. Volume 116, Issue 3, pp. 341-360.

Stoll, S., Abbot-Smith, K., Lieven, E. (2009). Lexically Restricted Utterances in Russian, German, and English Child-Directed Speech. *Cognitive Science* 33, 75–103.

Theakston, A. L., Lieven, E. V., Pine, J. M., Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *J. Child Lang*. 28(1):127-52.

Wang, H., Mintz, T. H. (2010). From Linear Sequences to Abstract Structures: Distributional Information in Infant-direct Speech. In Jane Chandlee, Katie Franich, Kate Iserman, Lauren Keil (eds.) *Proceedings Supplement of the 34th Boston University Conference on Language Development*.

Weisleder, A., Waxman, S. R. (2010). What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in Spanish and English. *J. Child Lang.* Nov; 37(5):1089-108. Epub. 2009 Aug 24.

Appendix

Training corpus data			Test file data		Results			
File ID	Group space (utterance types)	Number of words (types)	File ID	File size (utterance types)	Utterances mapped on some group		Novel utterances mapped on some group	
1a-2b	1000	525	3a	323	62	19%	19	6%
1a-5b	2741	912	6a	383	108	28%	29	7.6%
1a-8b	4476	1152	9a	393	112	28%	29	7.4%
1a-11b	6230	1379	12a	393	149	38%	29	7.4%
1a-14b	7800	1556	15a	369	123	33%	33	8.9%
1a-17b	9465	1747	18a	466	154	33%	45	9.6%
1a-20b	11209	1917	21a	395	162	41%	36	9.1%
1a-23b	12750	2066	24a	315	122	39%	23	7.3%
1a-26b	14321	2222	27a	369	130	35%	38	10.3%
1a-29b	15926	2367	30a	329	122	37%	33	10%
1a-32b	17260	2505	33a	284	92	32%	22	7.7%

Table 8. Details of the experiment.

Table 8 lists the major parameters of the experiment. File ID refers to the original file in the Anne corpus. For instance, *1a* refers to the text file obtained from *anne01a.xml*. The group space corresponds to the utterance types contained in the files identified. It also equals the number of agreement groups since each utterance type represents an agreement group. The test file represents the first 30 minutes of the next stage after the last training file. For example, the first row of the table says that the training corpus at that stage consisted of files *1a, 1b, 2a, 2b,* 1000 utterance types (agreement groups), 525 word types, and the test file was *3a* consisting of 323 non-identical utterances (types). As the results data further show, 62 of the 323 utterances (19%) were successfully mapped on at least one of the 1000 groups. Since 43 of the 62 utterances were already contained in the training corpus, the number of novel utterances mapped on at least one group was 19, which means about 6% of 323, the test file size.