LINGUISTIC COMPLEXITY AND FREQUENCY IN

AGRAMMATIC SPEECH PRODUCTION

Roelien Bastiaanse[1,2], Gosse Bouma[1] & Wendy Post[3]

[1]*Center for Language and Cognition Groningen (CLCG),*
University of Groningen, The Netherlands
[2]*University Medical Center Groningen (UMCG), The Netherlands*
[3]*Dept. of Epidemiology, University Medical Center Groningen (UMCG), The Netherlands*

Address for correspondence:
Prof. Dr. R. Bastiaanse
Dept. of Linguistics
University of Groningen
PO Box 716
9700 AS Groningen
The Netherlands

1

y.r.m.bastiaanse@rug.nl

Abstract

There is a long standing debate between aphasiologists on the essential factor that constitutes the behavioral patterns of loss and preservation in agrammatic Broca's aphasia. It has been suggested that linguistic complexity plays a crucial role: linguistically complex structures are more difficult to produce than linguistically simple ones. However, linguistic complex structures are often less frequent in a language; for example, simple active sentences are more frequently used than linguistically more complex passive sentences. Hence, it might be that it is not linguistic complexity but frequency that determines agrammatic behavior. Frequency may play a role at several levels. For agrammatic patients, for example, the frequency of sentence constructions may be crucial, whereas for fluent aphasic speakers word frequency influences performance.

The present study presents corpus frequency data for constructions that have previously been used to show the influence of linguistic complexity on Dutch agrammatic speech production. These are data on: (1) verb movement; (2) object scrambling; and (3) verbs with alternating transitivity. We compared the data of our corpus research with the performance of agrammatic speakers on the constructions. The conclusion is that frequency cannot account for the data, but linguistic complexity can. It is then discussed what 'linguistic complexity' exactly stands for, in terms of the word order deficit in agrammatic aphasia.

Introduction

Traditionally, Broca's aphasia has been described as a morphosyntactic disorder which caused the patients to speak agrammatically, meaning that function words and grammatical morphemes are omitted and/or substituted (Miceli et al., 1988; Saffran et al., 1989). It is a known fact, however, that the pattern of omission and substitution is language dependent (e.g. Grodzinsky, 1990) and that not all function words and grammatical morphemes are equally vulnerable (e.g. De Bleser et al., 1996; Bastiaanse et al., 2003). It has also been demonstrated that the error pattern largely depends on the structure of the sentence (e.g. Grodzinsky, 1995 for comprehension; Bastiaanse & Thompson, 2003 for production). Several theories have been formulated to account for the deficit at the sentence level in agrammatic aphasia. Some of them are representational accounts, such as the *Trace Deletion Hypothesis* (e.g. Grodzinsky, 1995) and the *Tree Pruning Hypothesis* (e.g. Friedmann, 2000). These assume that parts of the linguistic representations (traces, top of the syntactic tree) are gone due to brain damage. Other theories, such as the *Argument Structure Complexity Hypothesis* (Thompson, 2003) and the *Derived Order Problem Hypothesis* (Bastiaanse & Van Zonneveld, 2005), assume that the linguistic representations are intact, but due to processing disorders, some are harder to retrieve than others. What these theories have in common is that they hold linguistic complexity responsible for the problems that agrammatic patients encounter.

In any well-controlled study, sentences in different conditions are matched on length (to control for auditory memory) and the differences between the two conditions are minimalized. Therefore, a good test for comprehension of word order is one that compares, for example, subject and object relatives. The crucial difference in meaning between *the horse that chases the cow* and *the horse that the cow chases* is caused by the difference in word order. Agrammatic patients are known to perform poorer on object relatives than on subject relatives. Linguistic accounts use this as evidence for the influence of linguistic complexity, but one can argue that the poor performance on the object relatives may equally well be due to the low frequency of this construction in English. The concept that linguistically complex structures are of lower frequency is key for the following discussion.

Passives, notoriously difficult to comprehend (Grodzinsky, 1995) and produce (Bastiaanse & Edwards, 2004), are both linguistically more complex and of lower frequency than actives. One

should notice, however, that one of the characteristics of agrammatic speech is the omission of function words, which are highly frequent.

The purpose of the present study is to tear linguistic complexity and frequency apart. During the last decade, we published several studies on the influence of linguistic complexity on the performance on Dutch agrammatic speakers. The following linguistic topics were involved: verb movement (Bastiaanse & Van Zonneveld, 1998; Bastiaanse & Thompson, 2003), object scrambling (Bastiaanse et al., 2003) and verbs with alternating transitivity (Bastiaanse & Van Zonneveld, 2005). The results of these studies were interpreted as if the production of agrammatic speakers were influenced by linguistic complexity. This theory was captured in the *Derived Order Problem Hypothesis (DOP-H)* which assumes that every language has a base-order, for example, Subject – Verb – Object (SVO) for English and Subject – Object – Verb (SOV) for Dutch and German. All other word orders are derived. Sentences with derived word order are, according to the DOP-H, difficult to comprehend and produce for agrammatic patients. However, in the above mentioned results, frequencies of the tested constructions and the used verbs were not taken into account. For the present study we complement the data with corpus frequency counts. For all of the constructions and each individual verb that have been used in the experiments, corpus frequencies were established and it was calculated whether the data of the experiments on verb movement, object scrambling and verbs with alternating transitivity can be accounted for in terms of frequency.

Collecting frequency data

Three constructions have been used to study the effect of linguistic complexity in agrammatic aphasia: (1) verb movement; (2) object scrambling; (3) verbs with alternating transitivity. To obtain counts for syntactic constructions or for verbs occurring in a given syntactic environment, we need syntactically annotated data. We used two different corpora. The Corpus of Spoken Dutch (CGN) (Oostdijk, 2000)[1] is a 10 million word spoken language corpus, containing formal and informal monologues and dialogues. 1 million words have been annotated manually with syntactic dependency trees. The Twente News Corpus (TwNC)[2] is a 500 million words (25 million sentences) corpus consisting of materials from a number of Dutch national newspapers

---

[1] http://lands.let.kun.nl/cgn/
[2] www.vf.utwente.nl/~druid/TwNC/TwNC-main.html

covering the period 1994-2004. This corpus was parsed automatically using the Alpino parser (Bouma et al., 2001; van Noord, 2006). The output of the parser is a syntactic dependency tree which conforms to the annotation guidelines developed for CGN. Thus, the two corpora use identical syntactic annotation. Parse results are stored in XML, and we used XPath (a query language for XML-documents) to search for sentences that contained a given construction or a verb in a given context.[3]

We used the syntactically annotated part of CGN (1 million words, 100,000 sentences) and a small portion of the TwNC corpus (*Algemeen Dagblad*, January, 1999; 1.5 million words; 75,000 sentences) to collect frequency figures for a number of high-frequent syntactic constructions (i.e. the frequency of verbs in verb second and verb final position for various types of verbs, and the relative order of direct object and adverb). As these constructions are very frequent, and we were not interested in counts for specific lexical items, corpora of approximately 1 million words are sufficiently large to obtain reliable figures. The complete TwNC corpus was used to collect frequency data for specific verbs in a specific context. We first collected all sentences in the corpus containing (a root form of) the word, and then searched these sentences for occurrences of the verb in a given construction.

There are two important differences between the corpora we used. First, spoken and written language differ in many respects, and this may effect the frequency of syntactic constructions (see Roland et al., 2007, for a comparison of the frequency of a number of syntactic constructions in spoken and written English). Nevertheless, for those constructions that we investigated using both corpora, the resulting counts are quite similar. Second, CGN is manually annotated, whereas TwNC has been parsed automatically. The accuracy of the Alpino parser on newspaper text is quite high. Malouf & Van Noord (2004) report, for instance, that for newspaper texts, dependency relations are labeled with more than 91% accuracy. This is in line with state of the art statistical dependency parsers for English. Nevertheless, automatically annotated materials contain far more errors than manually corrected data. For the experiments involving verb position and the relative order of object and adverb, we manually inspected samples of the results, and concluded that the parser recognizes with high accuracy that a verb occurs in verb second or base position. The same is true for recognizing object-adverb and adverb-object order. The situation is rather different for recognizing verbs that can have both a transitive and unaccusative reading. Here we made a more thorough manual inspection of the data. The reason

---

[3] A technical description of the corpus and the tools we used can be found in Bouma & Kloosterman (2007).

is twofold. First, some of the verbs can occur without an object (*object drop*), without being unaccusative. These cases cannot automatically be distinguished completely from unaccusative readings. Second, many of the verbs used in the experiments are ambiguous and can occur as part of an idiomatic expression. As the corpus is only annotated for syntax, and not for word meaning, the correct readings of the verbs cannot be identified automatically. Therefore, frequencies had to be estimated on the basis of manual inspection of the automatically extracted samples. For all statistical comparisons, log transformed frequencies have been used, since these are more normally distributed and the effect of outliers is reduced. Also, the frequency of a word in a certain context has a logarithmic effect on, for example, processing time (Smith & Levy, 2008).

Study 1: Verb movement

Background

Dutch is a so-called *Subject – Object – Verb language*, meaning that the verb is base-generated in clause final position (see 1-3). In the matrix clause, however, the finite verb is 'moved' to second position, a rule that is called *Verb Second* (see 4).

(1)   *finite verb in embedded clause*
         de jongen die een boek <u>leest</u>
         the boy who a book <u>reads</u>

(2)   *infinitive*
         de jongen wil een boek <u>lezen</u>
         the boy wants a book to <u>read</u>

(3)   *participle*
         de jongen heeft een boek <u>gelezen</u>
         the boy has a book <u>read</u>

(4)   *finite verb in matrix clause*
         de jongen <u>leest</u>$_i$ een boek   i
         the boy <u>reads</u> a book

The 'i' in (4) indicates the original position of the verb (as shown by the co-indexation of the finite verb in Verb Second position). If it is assumed that derived order, as in (4), is linguistically more complex than base order, then the order Subject – Verb – Object in the matrix clause is more complex than the base Subject – Object – Verb order in the embedded clause. If, consequently, one assumes that linguistic complexity influences the performance of agrammatic speakers, then production of [finite verb – object] in the matrix clause is more difficult than [object - finite verb] in the embedded clause. This has been investigated by Bastiaanse & Thompson (2003). Agrammatic speakers were tested in two conditions: completing matrix clauses with (derived) [finite verb – object] and embedded clauses with (base) [object – finite verb]. The same pictures (and, thus, verbs and nouns) were used in both conditions. An example of both conditions is given in (5a-b).

(5a) *Matrix clause*

 Tester: 'Dit is de jongen die de tomaat snijdt en dit is de jongen die het brood

 snijdt. Dus deze jongen snijdt de tomaat en deze jongen …..'

 Patient: *'snijdt het brood'*

 (tester: 'This is the boy that the tomato cuts and this is the boy that the bread

 cuts. So, this boy cuts the tomato and this boy …..'

 Patient: '*cuts the bread'*)

(5b) *Embedded clause*

 Tester: 'Deze jongen snijdt de tomaat en deze jongen snijdt het brood. Dus dit

 is de jongen die de tomaat snijdt en dit is de jongen die …..'

 Patient: *'het brood snijdt'*

 (Tester: 'This boy cuts the tomato and this boy cuts the bread. So this is the boy

 that the tomato cuts and this is the boy that…'

 Patient: *'the bread cuts'*)

The results demonstrated that the matrix clause condition (43% correct) was more difficult than the embedded clause condition (73% correct) for the agrammatic speakers.

For the present study, several frequency comparisons have been made:

(1) Lexical verbs in base position (order OV; examples 1-3 above) versus lexical verbs in Verb Second position (order VO; example 4 above). This analysis was made to evaluate whether the analysis Dutch = SOV-language is correct.

(2) Finite verbs (including auxiliaries, copulas and modals) in base position (the embedded clause, example 1 above) versus finite verbs in Verb Second position (the matrix clause, examples 2-4 above). This analysis was performed to evaluate whether, in general, finite verbs are more often in base or in derived position.

(3) Finite *lexical* verbs in base position (example 1) versus Verb Second position (example 4 above). This was done to verify the findings of the second comparison for lexical verbs only, which were under study in Bastiaanse & Thompson, 2003)

(4) *Object* + finite lexical verbs in base position (example 1) versus finite lexical verbs in Verb Second position + *Object* (example 4). This comparison was made to find out whether finite transitive verbs (the constructions used in the experiment) are more often in base or in derived position.

(5) Finite verbs used in the experiment in base (example 1) versus Verb Second position (example 4). This was done to analyze whether the structures in which the verbs included in the test occur have the same frequency distribution as transitive verbs in general.

Results

Tables 1a+b give an overview of the counts and relative frequencies of constructions (1-4) mentioned above, both in CGN and in TwNC. Some comparisons are restricted to lexical verbs (+LEX), others include all verbs (lexical verbs plus auxiliaries, modals and copulas: -LEX). Some comparisons are for finite verbs only (+FIN), others cover all forms of a verb (-FIN). Finally, some counts are only for verbs selecting a direct object (+OBJ), others do not impose this restriction (-OBJ). The final two columns give the counts and relative frequencies of these verbs in Vfinal and Verb Second position. Appendix 1, Table A contains the frequencies of the individual verbs used in the experiments in the full TwNC and compares these with the agrammatic data.

[Table 1a+b about here]

The frequency analyses show the following results:

- *Lexical verbs (comparison 1).* The base order *object – lexical verb* is more frequent in Dutch than the derived order *lexical verb – object*. The difference is strongest in written language, where 60% of the lexical transitive verbs occurs in base position (i.e. in the order object -verb). Note that this can be seen as a consequence of the fact that Dutch is an SOV-language: all non-finite uses of a verb and some finite uses (embedded clause) give rise to the order OV, and only finite verbs in the matrix clause give rise to the order VO. The difference between spoken and written language may be attributed to the fact that the average sentence length in CGN is much shorter than in TwNC (approx 9 words per utterance in CGN and almost 20 words per sentence in TwNC), and thus the possibilities for using the base word order (which is restricted to nonfinite verbs and embedded clauses) are also smaller.

- *Finite verbs (comparisons 2, 3, and 4).* If we restrict attention to finite verbs, around 80% (CGN) and 73% (TwNC) of the verbs occurs in verb second position. This is true for finite verbs in general, for lexical finite verbs, and for lexical, transitive, finite verbs. This is, of course, not surprising: it shows that the matrix clause is four times as frequent as the (finite) embedded clause in spoken language and about three times as frequent in written language.

- *Finite verbs used in the experiment in base versus Verb Second position.* The occurrence in Verb Second position of the individual verbs used in the experiment (75.5%) is close to that found for all finite verbs together in the TwNC fragment. Although there is some variation among individual verbs, each verb in the test is found more often in Verb Second than in base position.

The results so far have been summarized in Figure 1.

[Figure 1 about here]

These analyses show that the theory that Dutch is an SOV-language is justified: lexical verbs are more often used in base position. However, finite verbs, the verb forms that were tested in the experiment, occur more often in Verb Second position, the position that is supposed to be derived and therefore linguistically complex. This holds for verbs in general (including modal verbs, auxiliaries and copulas), for lexical verbs and for transitive lexical verbs.

The graphical representation of the verbs used in the experiment and the performance of the agrammatic speakers is given in Figure 2. The frequencies of the individual verbs and performances of the agrammatic patients per verb are given in Appendix 1, Table B. A graphical representation is provided in Appendix 1, Figures A and B.

[Figure 2 about here]

To evaluate the effect of position and frequency of the individual verbs used in the experiment on agrammatic behavior, multilevel logistic regression analysis has been performed to adjust for the correlation within individuals. Predictors in this model were *condition* (Verb Second vs base position) and *frequency* (log transformed frequencies were used). MLWIN2.02 was used for this analysis. The results of this analysis were that the probability of a correct response was significantly smaller for the Verb Second position ($p<0.01$). There was no effect of frequency.

Conclusion

All the frequency data together show that the results on the experiment are not likely to be caused by the pattern of distribution of finite (lexical) verbs over base and Verb Second position. Linguistic complexity, however, can account for the behavior of the agrammatic speakers. The DOP-H rightly predicts that finite verbs in base position are easier to produce than finite verbs in derived, that is, Verb Second position.

Study 2: Object scrambling

Background

In a Dutch sentence, the order of the direct object and the adverbial phrase is optional: the adverbial phrase may either precede or follow the object. In the underlying form, the object is assumed to be the sister of the verb, and therefore adjacent to the verb. The [object – verb] complex can be preceded by an adverbial phrase (see 6; the examples are given in the embedded clause to avoid interference with verb movement and in accordance to the experiment). The object and adverbial phrase may change position; this is called *object scrambling*. Object

10

scrambling does not change the meaning of the sentence; it is only a pragmatic shift. In (7) an example is given.

(6)   dit is de jongen die <u>vandaag</u>   <u>het boek</u>   leest
        this is the boy who <u>today</u>   <u>the book</u>   reads

(7)   dit is de jongen die <u>het boek_i</u>   <u>vandaag</u>   i   leest
        this is the boy who <u>the book</u>   <u>today</u>   reads

The linguistic operation *object scrambling* is shown by the trace (i) in base position and the coindexation of the object in its derived position.

There are a few restrictions to object scrambling in Dutch: (unspecific) indefinite objects cannot scramble (8); personal pronouns in object position scramble obligatorily (9); when there is negation, scrambling is either obligatory (10) or changes the meaning of the sentence and thus cannot be considered truly optional

(8)   Indefinite NP ⇒ no scrambling

        base order:

        dit is de jongen die   <u>vandaag</u>   <u>een boek</u>   leest
        this is the boy who   <u>today</u>   <u>a book</u>   reads

        derived order:

        dit is de jongen die   *<u>een boek</u>   <u>vandaag</u>   leest
        this is the boy who   <u>a book</u>   <u>today</u>   reads

(9)   Pronouns ⇒ obligatory scrambling

        base order:

        dit is het meisje dat   *<u>vandaag</u>   <u>hem</u>   kust
        this is the girl who   <u>today</u>   <u>him</u>   kisses

        derived order:

        dit is het meisje dat   <u>hem</u>   <u>vandaag</u>   kust
        this is the girl who   <u>him</u>   <u>today</u>   kisses

(10)   *Negation ⇒ obligatory scrambling*

base order:

dit is de jongen die	*niet	sommige_boeken	leest
this is the boy who	not	some book	reads

derived order:

dit is de jongen die	sommige boeken	niet	leest
this is the boy who	some books	not	reads

In sentences without negation and with a definite object (that is not a personal pronoun), scrambling is optional. This is the sentence type that was used by Bastiaanse et al. (2003). Object scrambling is always possible in these sentences, but usually occurs when the object contains *given* information and the adverb is *new information*, following the pragmatic principle that given information precedes new information. In the first experiment, the object was varied and the adverb was constant (examples are given in 11a-b). Hence, in the scrambled condition (11b), the new information preceded given information, meaning that this condition was marked from both a syntactic and a pragmatic point of view.

(11a)   *Base order (given – new)*

Tester: 'Deze man snijdt de tomaat en deze man snijdt het brood. Dus dit is de man die vandaag de tomaat snijdt en dit is de man die ….'

Patient: *'vandaag het brood snijdt'*

Tester: 'This man cuts the tomato and this man cuts the bread. So, this is the man who today the tomato cuts and this is the man who ….'

Patient: *'today the bread cuts'*

(11b)   *Scrambled order (new - given)*

Tester: 'Deze man snijdt de tomaat en deze man snijdt het brood. Dus dit is de man die de tomaat vandaag snijdt en dit is de man die ….'

Patient: *'het brood vandaag snijdt'*

Tester: 'This man cuts the tomato and this man cuts the bread. So, this is the man who the tomato today cuts and this is the man who ….'

Patient: *' today the bread cuts'*

In the second experiment (see 12a-b), the object was held constant and the adverb varied. This implies that in the scrambled condition, the given information (the object) preceded the new information (the adverb) and this construction is only syntactically marked.

(12a)   *Base order (new – given)*

Tester: 'Deze man leest liggend en deze man leest zittend. Dus dit is de man die liggend het boek leest en dit is de man die ….'

Patient: *'zittend het boek leest'*

Tester: 'This man reads lying and this man reads sitting. So, this is the man who lying the book reads and this is the man who ….'

Patient: *'sitting the book reads'*

(12b)   *Scrambled order (given – new)*

Tester: 'Deze man leest liggend en deze man leest zittend. Dus dit is de man die het boek liggend leest en dit is de man die ….'

Patient: *'het boek zittend leest'*

Tester: 'This man reads lying and this man reads sitting. So, this is the man who the book lying reads and this is the man who ….'

Patient: *'the book sitting reads'*

There was no significant difference between the performance on the base order condition in both experiments, nor in the scrambled order conditions (MWU; base order z=- 0.806, p=0.42; scrambled order z=-1.59; p=0.11), meaning that the order of given and new information did not play a critical role. However, the agrammatic speakers were significantly more impaired in the scrambled conditions on both tests. In scrambled sentences, the word order is derived, and therefore Bastiaanse et al. concluded that derived order is difficult for agrammatic speakers.

Alternatively, however, object scrambling may be a rare phenomenon, resulting in poor performance of agrammatic speakers. In order to evaluate this possibility, two analyses were performed[4]:

(1) The order *adverbial phrase – object* versus *object – adverbial phrase*. This comparison reveals which order is more frequent in general.

(2) The order adverbial phrase – object with definite article versus object with definite article – adverbial phrase. These were the constructions that have been tested.

Results

The following frequencies were found:

(1) The order adverbial phrase – object versus object – adverbial phrase. In sentences with an object and an adjacent adverbial phrase (where scrambling may be either obligatory, optional, or excluded), scrambling is applied 46% of the time in spoken language and 40% of the time in written language.

(2) The order adverbial phrase – object with definite article versus object with definite article – adverbial phrase. In those cases where the object is definite, and scrambling is optional, scrambling is applied 41% of the time in spoken language, and 46% of the time in written language.

In Figure 3, the data are summarized. In Appendix 2, the exact frequencies (Table A: all NPs; Table B: definite NPs) and the agrammatic data (Table C) are given.

[Figure 3 about here]

Conclusion

These frequency data show that scrambling is not a rare phenomenon in Dutch: almost half of the times the order of the adverbial phrase and the object is scrambled. However, agrammatic speakers have serious problems with scrambled objects. Overall, on the two tests, they produce a correctly scrambled sentence in less than 20% of the items, significantly less often than a correct sentence in base order (almost 75% correct). Interestingly, the most frequent error in the scrambled condition is 'descrambling'. This means that a frequency account is not the most

---

[4] For these analyses sentences with negation were excluded in order not to overestimate the frequency of the scrambled order.

likely explanation for the agrammatic data. Again, the DOP-H, based on linguistic theory predicts the data correctly.

Study 3: verbs with alternating transitivity

Background

In Dutch, like in English, some verbs can have both a transitive and an unaccusative reading (13, 14). According to linguistic theories (e.g. Burzio, 1986; Levin & Rappaport-Hovav, 1995) the unaccusative is derived from the transitive sentence. In an unaccusative sentence, there is no agent role to fulfill the grammatical role of subject, therefore, the theme of the sentence is 'moved' to subject position (see 15), as demonstrated by the coindexation of the trace (i) and the theme.

(13)   *transitive*

    the boy breaks the glass

(14)   *unaccusative*

    the glass breaks

(15)   *derivation*

    the glass$_{(i\ theme,\ subject)}$ breaks   i

The theory that the order in the unaccusative condition is more complex than in the transitive condition was reason for Bastiaanse & Van Zonneveld (2005) to study agrammatic performance on both sentence types. They also included fluent aphasic speakers in this experiment. The data showed that agrammatic speakers have more problems constructing an unaccusative sentence than a transitive sentence, with the same verb. This difference was not found for the fluent aphasic speakers. However, it is well conceivable that the unaccusative construction is less frequent than the transitive construction. To test this hypothesis, we are interested in frequency data for the verbs used in the experiment, both in their transitive and unaccusative use. As first approximation, one might count the number of finite transitive and intransitive uses of these verbs in the corpus. We restrict our attention to finite verbs, in order to avoid sentences where the verb is governed by the verb *laten (to cause)*, as these generally introduce an ambiguity between

15

the transitive and unaccusative use of the verb. Unfortunately, these counts are still very muddy. There are two reasons for this. First, a number of verbs (such as 'to smoke' in 16a-c) allow both an unaccusative as well as an unergative (object drop) reading.

(16a)   Kim smokes a cigarette

(16b)   The fire is smoking

(16c)   Kim smokes regularly

Both (16b) and (16c) give rise to a syntactic dependency structure in which the verb has a subject and no direct object. Note that the dependency structures produced by the parser are purely syntactic, and do not encode thematic roles or semantic distinctions.

Second, almost all verbs used in the experiment have readings and uses without alternating transitivity, for example, the verb *luiden* (*to ring*). This verb has a transitive and unaccusative reading (the priest is ringing the bell, the bells are ringing, however, the most frequent use of this verb is with the meaning *to be (called),* as in *het antwoord luidt 'nee'*: 'the answer is 'no''. As the corpus does not distinguish between word senses, this use of the verb cannot be distinguished from the other uses. We wanted to exclude these cases from our frequency counts, as they have a meaning that is different from the meaning used in the experiments with aphasic speakers, and also it does not allow for both an unaccusative and transitive use.

We therefore estimated the frequency of the transitive and unaccusative use of the verbs used in the experiments by manually inspecting, for each verb, 100 sentences where the verb was used transitively and 100 sentences where the verb was used intransitively. We counted the number of true positives in these samples. On the basis of these counts, we estimate the number of true positives in all extracted sentences, and then compute the relative frequencies. The results are given in Appendix 3, Table A. Note that the number of true positives per 100 extracted transitive or intransitive sentences varies greatly. As explained above, this should not be interpreted as a weakness of the parser, but it simply means that many of the verbs used in the experiment are ambiguous in meaning. In 11 out of 28 samples (5 of the intransitive uses, 6 of the transitive uses), the number of true positives is less than 50%. An extreme case is *roken (to smoke)*, for which no true positives were found (i.e. the dominant intransitive reading is *he/she smokes*, which is an instance of object drop).

Results

In Appendix 3, the exact and relative frequencies (Table A, Figure A) and aphasic data (Table B, Figure B) are given. Overall, the verbs were used about as frequently in unaccusative (51%) as in transitive (49%) condition.

All verbs but one ('drowning') are used more often correctly in a transitive sentence by the agrammatic speakers. The most frequent error sentence in the unaccusative condition for the agrammatic speakers is producing the transitive construction, that is, introducing a person (often 'I') that is not in the picture; e.g. the book is burning → I am burning the book). The fluent aphasic speakers produce 49% of the transitive construction correctly and 46% of the unaccusative constructions.

Like in the first experiment, multilevel logistic regression analysis has been performed to adjust for the correlation within individuals. Predictors in this model were *group* (agrammatic vs fluent aphasia), *condition* (unaccusative vs transitive) and *frequency* (log transformed frequencies were used). There was no effect of frequency. The interaction effect of group times condition was significant ($p<0.01$). Agrammatic speakers had a larger probability of correct response in the transitive condition. Fluent aphasic speakers had a smaller probability of correct response in the transitive condition.

In Figure 4 the frequencies and the aphasic performances for the unaccusative and transitive condition are graphically represented.

[Figure 4 about here]

Conclusion

Again, the agrammatic behavior cannot be explained by the frequencies of the constructions tested. Linguistic complexity seems to be a more reliable predictor. The DOP-H, which is developed to account for the word order deficits in agrammatism, rightly predicts the results: the derived order is difficult to produce for agrammatic speakers.

Discussion

17

These data demonstrate that the problems of agrammatic speakers are not caused by the frequency with which structures are used in a language. Highly common structures are difficult as well, if they are syntactically complex. However, some objections can be made to the interpretations of the data of each of the experiments and the related conclusions above. In the next sections, alternative explanations will be discussed. It will be shown that although frequency and other theories might explain part of the data presented here, no theory, except one based on linguistic complexity, can cover all the data. It is important to note that we do not claim that linguistic complexity is the only factor, not that it is a matter of complex versus simple. It is very likely that linguistic complexity is layered and that the number of syntactic operations and/or the length of these operations (e.g. scrambling versus movement) have different effects on agrammatic behavior. Also, it might very well be the case that there is an interaction between frequency and linguistic complexity, which was not caught by the present data. There is evidence from psycholinguistic and modeling studies that there are many factors that contribute to overall complexity, including frequency of constructions and lexical items and linguistic complexity (Chang et al., 2006).

Still, the results of the three meta-analyses show that the frequency of neither the tested constructions nor the used verbs can account for the data. (1) Finite verbs are more often in derived than in base position. This not only holds for finite verbs in general, but also for finite lexical verbs, and for the verbs used in the experiment. This is opposite to the pattern found in the agrammatic data: finite verbs in derived position are difficult to produce compared to the finite verbs in base position. (2) Object scrambling occurs in approximately half of the constructions containing an adjacent object and adverbial phrase. Nevertheless, producing sentences with scrambled objects is extremely difficult for agrammatic speakers. (3) Verbs with alternating transitivity are as frequently used in unaccusative (=derived) transitive (base order) sentences, but the unaccusative construction is, again, difficult for agrammatic speakers. We will finish this section with some thoughts of the nature of concepts like 'derived order' and 'movement'.

Verb movement

In Dutch, finite lexical verbs are used more often in Verb Second than in base position (75-80% versus 25-20%, respectively). This was reason to reject a frequency account for the production of finite verbs in these to positions of the agrammatic speakers (43% correct in Verb Second

position versus 73% correct in base position). The overall frequencies of lexical verbs (including the finite verbs, and the infinitives and participles) in base and Verb Second position show that lexical verbs are most often in clause final position (60% versus 40%), thus following the pattern of performance of the agrammatic speakers.

The question is whether the frequency of verb *position* is crucial, or the frequency of the *structure*. A frequency explanation can only be accepted when verb *position* is considered most important. In that case different verb forms are taken together: finite verbs, infinitives, participles. From a linguistic point of view that is not sensible. If sentence structure is supposed to be essential, then a frequency account cannot explain the agrammatic data, but a linguistic complexity account can.

Object scrambling

A similar reasoning cannot be applied to the object scrambling data. Although object scrambling in general occurs slightly less often than non-scrambling, there is hardly a difference between the two conditions. This difference is even smaller when only sentences with definite objects are taken into account, which are the structures that were tested. Notice that the frequency counts are as conservative as possible: were sentences with negation included, that often scramble obligatory, then the frequency of scrambled and unscrambled definite objects is practically equal (850 scrambled vs. 846 unscrambled word orders in the TwNC fragment). However, whatever way the frequencies are used, the agrammatic data cannot be explained by the frequency of scrambled and unscrambled constructions. A linguistic theory that includes object scrambling as an operation predicts the performance perfectly.

Verbs with alternating transitivity

The data of the experiment with verbs with alternating transitivity show no relation between frequency and agrammatic performance: there is a low and insignificant correlation. Another explanation of the results of this study has been discussed extensively in the original paper (Bastiaanse & Van Zonneveld, 2005). This relates to the animacy of the subject in the transitive condition. A reason to reject this alternative is that there is no theory that can explain why patients with a *grammatical* deficit should have problems with inanimate subjects and patients with a *lexical* deficit (the fluent aphasic speakers in the experiment) not.

The nature of Derived Order and Syntactic Movement

Although the above discussion shows that frequency of the used constructions or of the used verbs cannot account for the data observed in the aphasia experiments, there are a few alternative explanations, varying from alternative frequency counts to intuitions. These explanations may all account for some of the data, but not for the data of all three experiments. Alternatively, the Derived Order Problem Hypothesis covers all the data. This is not surprising, since this hypothesis is based on the results of the experiments. It should be kept in mind, though, that all three experiments have been developed from a linguistic point of view: we aimed to test whether syntactic movement, as assumed in the current syntactic theories, predicts agrammatic behavior. Hence, theories on verb movement, object scrambling and derived subjects underlie each of the experiments and they describe the nature of the deficit very well.

Terms like 'syntactic movement' and 'derived order' may need some clarification. What is 'syntactic movement' and when does it result in 'derived order'? Chomsky (1957) introduced concepts like 'transformation' and 'syntactic movement' to describe the possibility of different word orders in a language. At a later stage, the distinction between 'covert' and 'overt' movement was made (see Chomsky, 1995). The term 'covert movement' refers to movement that is not visible, that is, movement that doesn't change word order. This is, in our view, a theoretical construct that is needed to explain some linguistic phenomena, but it is not useful to describe the deficits observed in speech impairments. 'Overt movement' results in word order changes, that is, the resulting word order is different than the base order in a given language. So, a language has a base order and through 'overt movement' other word orders may be 'derived'. This means that constituents that naturally belong to each other are separated. In generative linguistics, it is assumed that a sentence is built by combining two constituents, a head and a complement, with another constituent. For instance, a verb (*to read*) is combined with an internal argument (e.g. an object *book*). This so-called 'head-complement' combination is then combined with a third constituent, for example, an external argument (e.g. a subject *man*) or an adjunct (e.g. *yesterday*). This combining of constituents by adding a new phrase to a head-complement combination is recursive and is often represented by syntactic trees or between square brackets. The idea of the Derived Order Problem Hypothesis is that structures in which the head-complement structures are in their natural order (e.g. Subject – Verb – Object for English; Subject – Object – Verb for German and Dutch) are linguistically simple and structures where the head and its complementing constituent are separated, that is, sentences in derived order, are complex. The

results of the experiments show that when complements and heads are separated, the structure is difficult for agrammatic speakers. In Dutch the object is followed by the verb. When the object – verb chain is broken, as in Verb Second constructions in the matrix clause, the sentence becomes more complex and is difficult to produce for agrammatic speakers. The same holds when an adjunct interrupts the object – verb chain in scrambled sentences and when the internal argument is in subject position, as in unaccusative sentences. Although sentences in derived order are usually lower in frequency than base order sentences (e.g. passives are less frequent than actives), this is not always true as shown by the analyses of the present study.

Linguistic complexity is not only expressed through word order, there are other kinds of complexity as well. The findings are interesting with respect to argument structure. Thompson's *Argument Structure Complexity Hypothesis (ASCH*: Thompson, 2003) predicts that the more complex argument structure of a verb is, the more difficult this verb is for agrammatic speakers. This is an alternative account for the data of the experiment on verbs with alternating transitivity, although the idea is the same: what is linguistically complex is difficult for agrammatic speakers. However, Thompson also found that verbs with no internal argument (which she calls 'one-place verbs') are used more often in agrammatic spontaneous speech than verbs with one internal argument (that she calls 'two-place verbs'), whereas the opposite is observed in non-brain-damaged speech (Thompson et al., 1996). A similar phenomenon has been described for Dutch (Bastiaanse & Jonkers, 1998) and Italian (Rossi & Bastiaanse, 2008): in the agrammatic speech samples sentences without an internal argument are overused, whereas this sentence type is not the most frequent one in normal speech. This demonstrates once more that frequency is not the main factor.

References

Bastiaanse, R. & Edwards, S. (2004) Word order and finiteness in Dutch and English Broca's and Wernicke's aphasia. *Brain and Language, 89*, 91-107.

Bastiaanse R. & Jonkers R. (1998). Verb retrieval in action naming and spontaneous speech in agrammatic and anomic aphasia. *Aphasiology*, *12*, 951-969.

Bastiaanse, R., Hugen, J., Kos, M. & Zonneveld, R. van (2002) Lexical, morphological and syntactic aspects of verb production in agrammatic aphasics. *Brain and Language, 80,* 142-159.

Bastiaanse, R., Koekkoek J. & Van Zonneveld, R. (2003) Object scrambling in Dutch Broca's aphasia. *Brain and Language, 86,* 287-299.

Bastiaanse, R, Jonkers, R., Ruigendijk, E. & Zonneveld, R. van (2003) Gender and case in agrammatic production. Cortex, 39, 405-417.

Bastiaanse, R. & Thompson, C.K. (2003) Verb and auxiliary movement in agrammatic Broca's aphasia. *Brain and Language, 84,* 286-305.

Bastiaanse, R. & Zonneveld, R. van (2005) Sentence production with verbs of alternating transitivity in agrammatic Broca's aphasia, *Journal of Neurolinguistics, 18*, 57-66.

Bouma, G., Noord, G. van & Malouf, R. (2001) Alpino: wide-coverage computational analysis of Dutch. In Daelemans, W., Simaan, K., Veenstra, J. & Zavrel, J. (eds.) *Computational Linguistics in The Netherlands 2000* (pp45-59). Amsterdam: Rodopi.

Bouma, G. & Kloosterman, G. (2007) Mining syntactic annotated corpora with Xquery. In Boguraev, B. et al. (eds.) *Proceedings of the Linguistic Annotation Workshop* (pp. 17-24). Prague: Association for Computational Linguistics.

Burzio, L. (1986) *Italian Syntax*. Dordrecht: Reidel.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*, 234-272.

Chomsky, N. (1957) *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1995) *The Minimalist Program*. Cambridge MA: MIT Press.

De Bleser, R., Bayer, J. & Luzzatti, C. (1996) Linguistic theory and morphosyntactic impairments in German and Italian aphasics. *Journal of Neurolinguistics, 9,* 175-185.

Friedmann, N. (2000) Moving verbs in agrammatic production. In Bastiaanse, R. & Grodzinsky, Y. (eds.) *Grammatical Disorders in Aphasia: A Neurolinguistic Perspective*. (pp. 152-170). London: Whurr.

Grodzinsky, Y. 1990. *Theoretical Perspectives on Language Deficits*. Cambridge, MA: MIT Press.

Grodzinsky, Y. 1995. A restrictive theory of agrammatic comprehension. *Brain & Language, 51*, 26-51.

Levin, B. & Rappaport-Hovav, M. (1995) *The Unaccusative Hypothesis at the Syntax-Semantics Interface.* Cambridge (MA): MIT-Press.

Malouf, R & Noord, G. van (2004) Wide coverage parsing with stochastic attribute-value grammars. IJCNLP Workshop *Beyond Shallow Analyses – Formalisms and Statistical Modeling for Deep Analyses.* Hainan

Miceli, G., Silveri, M., Romani, C., & Caramazza, A. (1989) Variations in the pattern of omissions and subsititutions of grammatical morphemes in the spontaneous speech of so-called agrammatic patients. *Brain and Language, 36,* 447-492.

van Noord, G. (2006) At Last Parsing is Now Operational, in P. Mertens, C. Fairon, A. Dister, and P. Watrin (eds.), *Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles* (TALN6), pp. 20-42.

Oostdijk, N. (2000) Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde* 5, 280–284.

Roland, D., Dick, F., & J. Elman (2007) Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57, 348-379.

Rossi, E. & Bastiaanse, R. (2008) Spontaneous speech in Italian agrammatic aphasia: A focus on verb production. *Aphasiology 22,* 347-362.

Saffran, E.M., Berndt, R. & Schwartz, M.F. (1989) The quantitative analysis of agrammatic production. Procedures and data. *Brain and Language, 37*, 440-479.

Smith. N.J. & Levy, R. (2008) Optimal processing times in reading: A formal model and empirical investigation. Paper presented at *30th Annual Meeting of the Cognitive Science Society,* June 23-26, Washington DC, USA.
http://idiom.ucsd.edu/~rlevy/papers/smith-levy-2008-cogsci.pdf

Thompson, C.K. (2003) Unaccusative verb production in agrammatic aphasia: The argument structure complexity hypothesis. *Journal of Neurolinguistics, 16*, 151-167.

Thompson, C.K., Shapiro, L.P., & Schendel, L. (1995). Analysis of verbs and verb-argument structure. A method for quantification of aphasic language production. *Clinical Aphasiology*, *23*, 121-140.

Appendix 1: Verb movement experiment

Table A: Raw and relative frequencies (in the full TwNC) of the finite lexical verbs used in the experiment in Verb Second (VO).

		$V_{fin}O$-raw	$V_{fin}O$-%	OV_{fin}-raw	OV_{fin}-%
aaien	*to pat*	214	71.8	84	28.2
bekijken	*to watch*	4,642	62.6	2,800	37.4
breien	*to knit*	94	70.5	39	29.5
drinken	*to drink*	4,759	74.5	1,621	25.5
eten	*to eat*	6,909	74.2	2,361	25.8
filmen	*to film*	1,061	79.0	281	21.0
gooien	*to throw*	8,826	74.6	2,987	25.4
groeten	*to greet*	696	85.0	123	15.0
krabben	*to scratch*	295	77.7	85	22.3
lezen	*to read*	11,199	65.5	5,754	34.5
omhelzen	*to hug*	738	70.9	295	29.1
snijden	*to cut*	3,606	85.5	614	14.5
tekenen	*to draw*	7,247	84.1	1,342	15.9
volgen	*to follow*	21,132	75.6	6,724	24.4
zien	*to see*	105,458	80.9	24,002	19.1

Table B: Percentage of occurrence in Verb Second and base order and the number of agrammatic speakers (n=9) that constructed a grammatical sentence in Verb Second ($V_{fin}O$) and base (OV_{fin}) order per verb. The mean of the percentages is the macro average (i.e. the average of the percentage per verb). The micro averages (obtained by summing the VO and OV counts) are 77.8 (VO) and 22.2 (OV).

		finite form		agrammatic performance	
		$V_{fin}O$	OV_{fin}	$V_{fin}O$	OV_{fin}
aaien	to pat	<u>71.8</u>	<u>28.2</u>	3	8
bekijken	to watch	<u>62.6</u>	<u>37.4</u>	4	7
breien	to knit	<u>70.5</u>	<u>29.5</u>	5	7
drinken	to drink	<u>74.5</u>	<u>25.5</u>	5	7
eten	to eat	<u>74.2</u>	<u>25.8</u>	2	7
filmen	to film	<u>79.0</u>	<u>21.0</u>	5	6
gooien	to throw	<u>74.6</u>	<u>25.4</u>	5	6
groeten	to greet	<u>85.0</u>	<u>15.0</u>	6	6
krabben	to scratch	77.7	22.3	5	6
lezen	to read	<u>65.5</u>	<u>34.5</u>	1	6
omhelzen	to hug	70.9	29.1	5	7
snijden	to cut	<u>85.5</u>	<u>14.5</u>	3	6
tekenen	to draw	<u>84.1</u>	<u>15.9</u>	1	5
volgen	to follow	<u>75.6</u>	<u>24.4</u>	5	7
zien	to see	<u>80.9</u>	19.1	3	7
Mean		75.5	24.5	3.86	6.53

Appendix 1: Verb movement experiment –cont.

Figure A: Relative frequencies of the finite lexical verbs used in the experiment in Verb Second (VO) and base (OV) order on 500 million words.
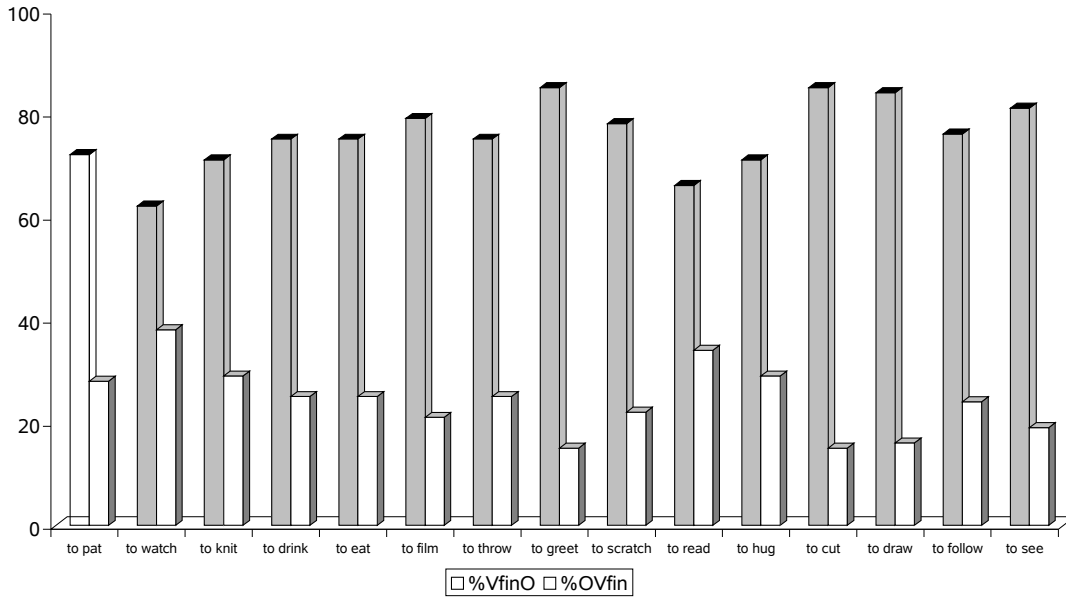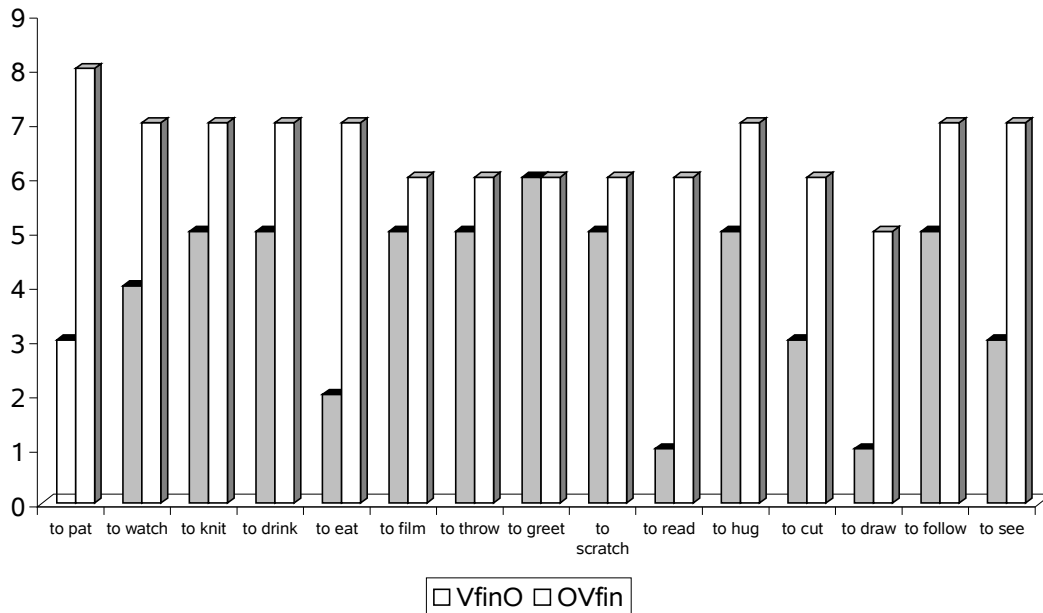

Figure B: Number of agrammatic speakers (n=9) who constructed a correct sentence in Verb Second ($V_{fin}O$) and base (OV_{fin}) order.

Appendix 2: Object Scrambling experiments

Table A: Comparison 1: Absolute frequencies of Adverbial Phrase + Object (base order) and Object + Adverbial Phrase (object scrambling) order in CGN and a section of TwNC (Algemeen Dagblad, January, 1999, 1.5 million words).

	Adv + Obj (base order)		Obj + Adv (scrambled order)	
	count	%	count	%
CGN	4,695	53.5	4,076	46.5
TwNC	4,648	60.1	3,091	39.9

Table B: Comparison 2: Absolute frequencies of Adverbial Phrase + <u>Definite</u> Object (base order) and <u>Definite</u> Object + Adverbial Phrase (object scrambling) order on 1.5 million words.

	Adv + DefObj (base order)		DefObj + Adv (scrambled order)	
	count	%	count	%
CGN	290	59.4	198	40.6
TwNC	751	54.3	631	45.7

Table C: Mean score correct of the agrammatic speakers on the Object Scrambling test 1 (n=8) and test 2 (n=6). Maximum score for test 1 is 16; maximum score for test 2 is 8 per condition.

	Base order	Scrambling
Test 1	10.75	5.00
Test 2	6.50	0.67

Appendix 3 Verbs with alternating transitivity experiment

Table A: Estimated relative frequency of 14 verbs used in their transitive and unaccusative sense in TwNC. Columns 3 and 4 give the counts for the intransitive and transitive use of the verb, columns 5 and 6 give the number of cases where the verb is used with the intended word sense in 100 random sentences from the extracted sample, and columns 7 and 8 give estimates for the relative frequency of the unaccusative and transitive use for the intended word sense.

		counts		est. correct		%	
		iv	tv	unacc.	trans.	unacc.	trans.
breken	*to break*	2,591	4,990	1,969	4,341	31.2	68.8
draaien	*to spin*	10,421	7,407	2,709	1,852	59.4	40.6
koken	*to cook/boil*	1,442	2,211	606	1,968	23.5	76.5
luiden	*to ring*	13,445	4,082	269	743	26.6	73.4
oplossen	*to solve*	417	15,140	296	204	1.6	98.4
rijden	*to drive*	19,069	11,428	6,102	1,485	80.4	19.6
rollen	*to roll*	3,646	738	3,500	244	93.5	6.5
roken	*to smoke*	0	2,114	0	2,114	0.0	100.0
scheuren	*to tear*	691	964	332	819	28.8	71.2
smelten	*to melt*	730	426	723	290	71.4	28.6
spelen	*to play*	38,969	78,982	1	9	5,2	94,8
stuiteren	*to bounce*	337	71	334	7	97.9	2.1
verbranden	*to burn*	510	663	479	623	43.5	56.5
verdrinken	*to drown*	1,426	149	1,397	80	94.6	5.4

Table B: Number of the agrammatic (n=8) and fluent aphasic (n=8) speakers who constructed a correct sentence in the transitive and unaccusative condition correctly.

		agrammatic speakers		fluent speakers	
		unaccusative	transitive	unaccusative	transitive
breken	*to break*	4	5	4	2
draaien	*to spin*	3	5	4	3
koken	*to boil/cook*	2	5	4	5
luiden	*to ring*	4	5	5	4
oplossen	*to solve*	1	4	3	2
rijden	*to drive/ride*	5	8	3	7
roken	*to smoke*	1	7	3	4
rollen	*to roll*	3	7	5	4
scheuren	*to tear*	0	7	1	4
smelten	*to melt*	3	6	5	3
spleen	*to play*	3	6	4	4
stuiteren	*to bounce*	3	4	4	3
verbranden	*to burn*	2	7	4	3
verdrinken	*to drown*	4	3	6	4
Mean		2.7	5.3	4.1	3.2

**Appendix 3 Verbs with alternating transitivity experiment –cont.**

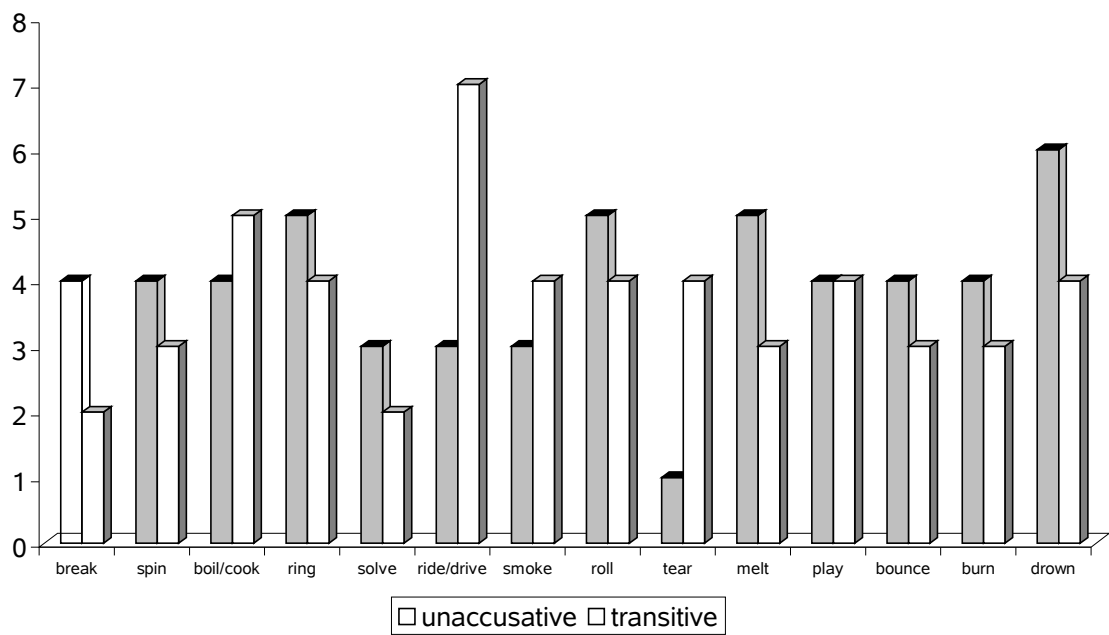*Figure A: The relative frequencies of the verbs in unaccusative and transitive condition*
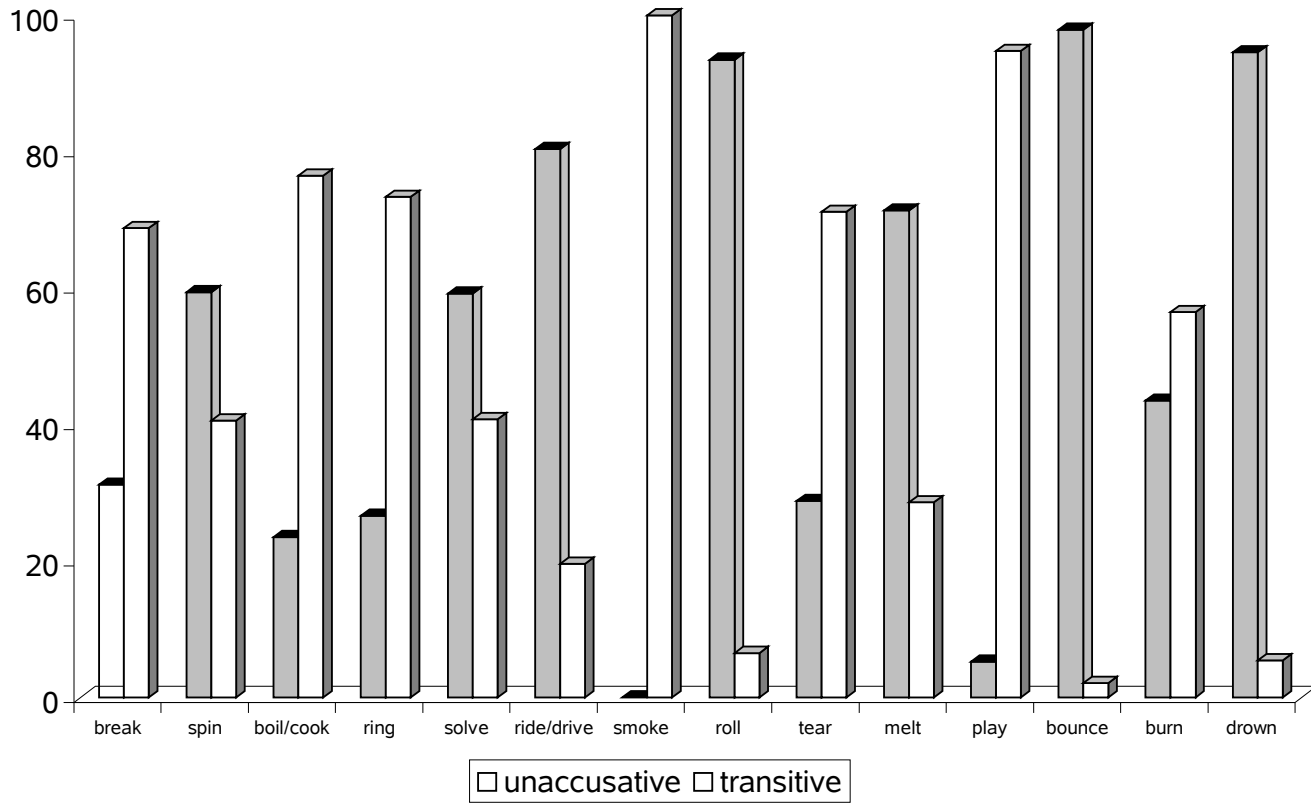
Appendix 3: verbs with alternating transitivity experiment –cont.

Figure B Number of times the individual verbs were used correctly in a sentence by the agrammatic (n=8; above) and fluent (n=8; under) speakers.
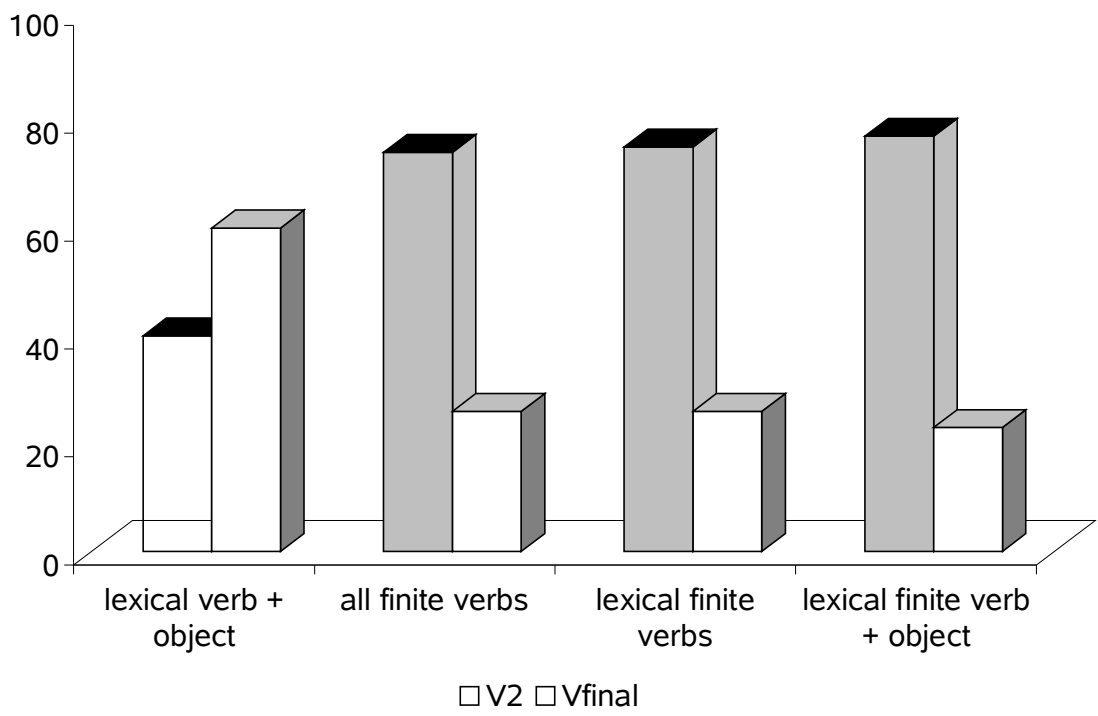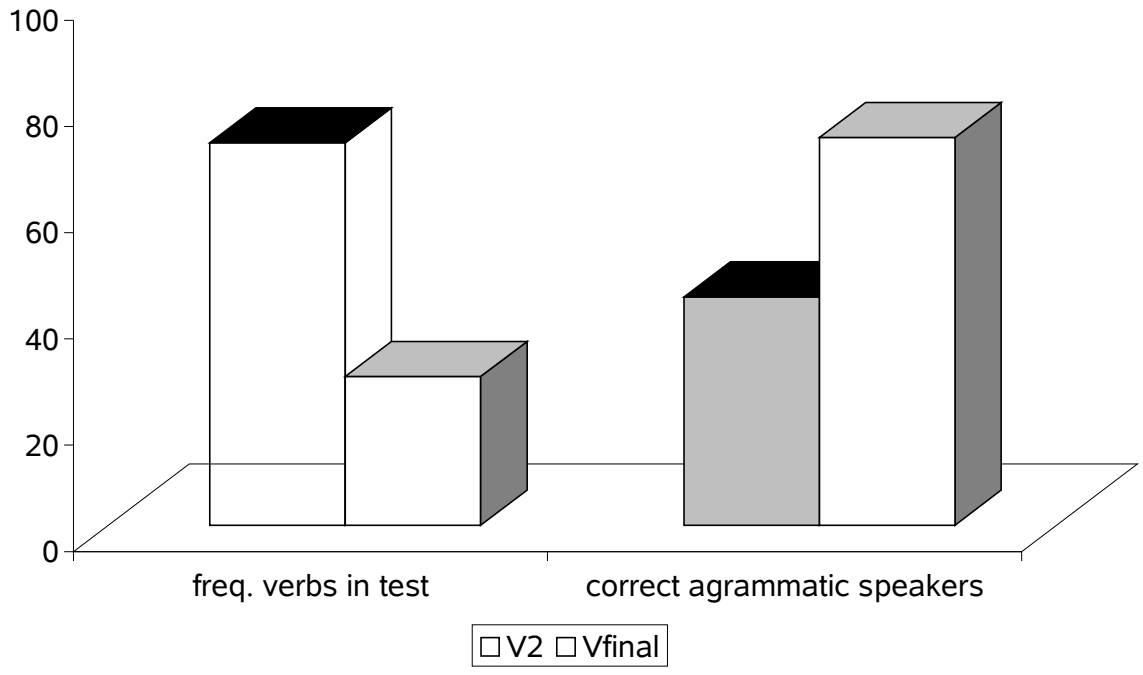
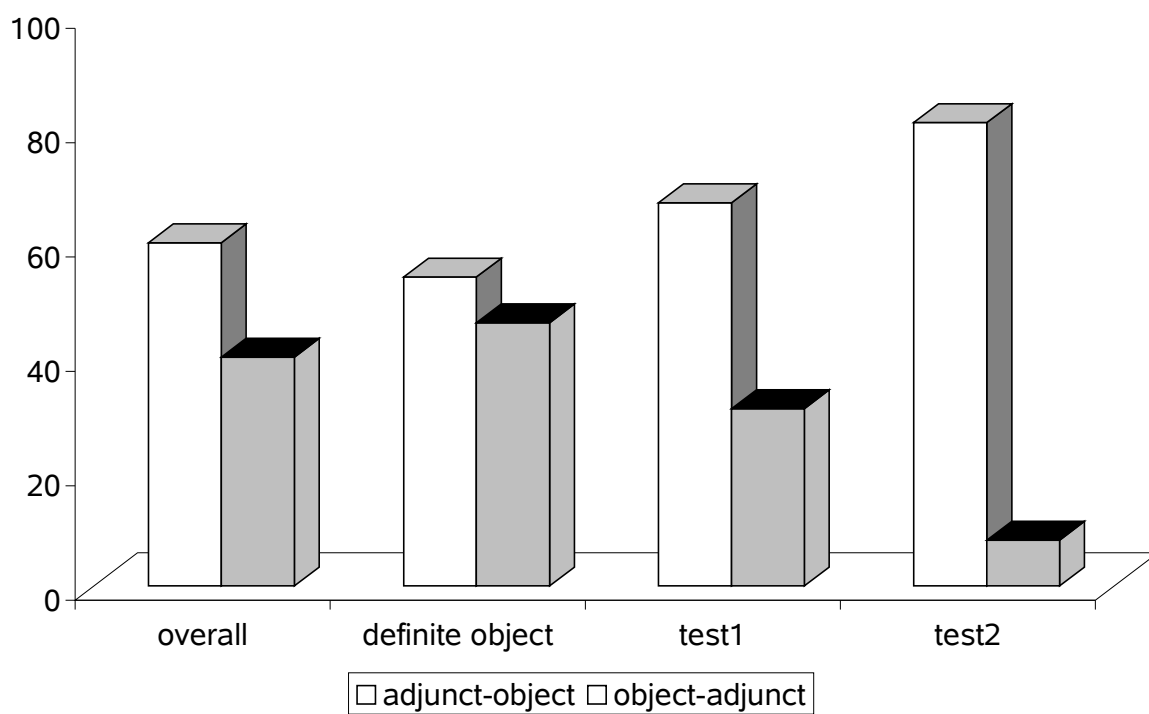## Agrammatic speakers
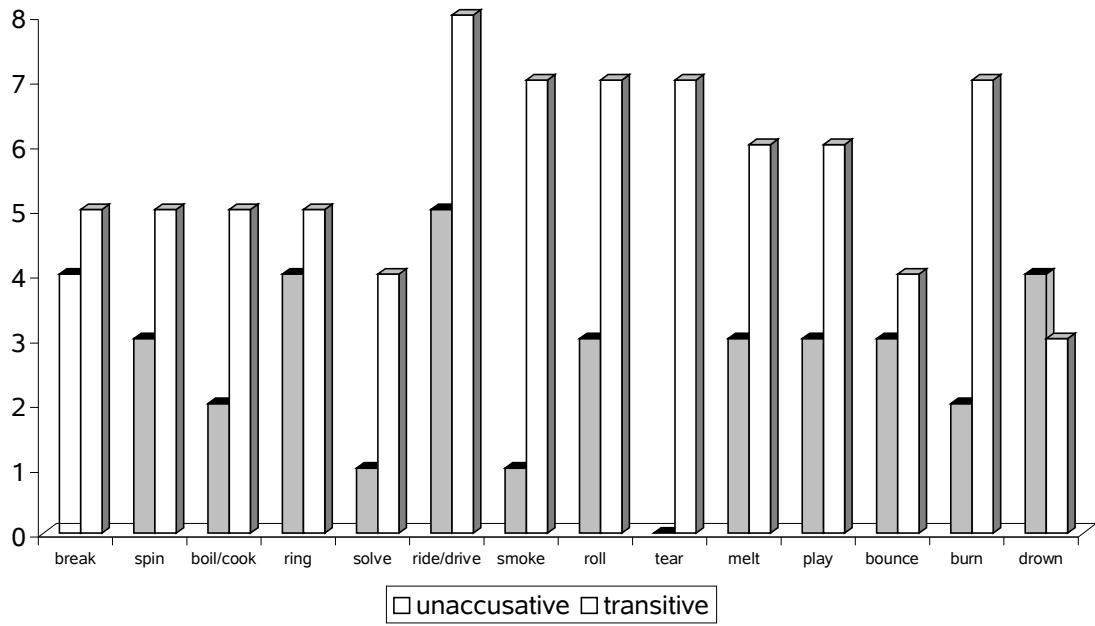
Fluent speakers

Legends to the figures

Figure 1   The relative frequencies in percentages of Dutch finite and nonfinite verbs in Verb Second (V2) and base (Vfinal) position.

Figure 2   The relative frequencies in percentages of finite verb – object (V2=derived) and object – finite verb (Vfinal=base) sequences of the verbs used in the experiment and the mean percentages correct of the agrammatic speakers.

Figure 3   The relative frequencies in percentages of adjunct – object (=base) and object – adjunct (derived) order. Most left bar are percentages for all of these sequences; second left bar for those with a definite object. The two right bars are the mean percentages correct of the agrammatic speakers on the two tests.

Figure 4   The relative frequencies in percentages of the transitive (=base order) and unaccusative (=derived order) use of the verbs used in the experiment (left bar). The two right bars represent the percentage correct of the agrammatic and the fluent speakers.

.

100

80

60

40

20

0

lexical verb +
object

all finite verbs

lexical finite
verbs

lexical finite verb
+ object

□ V2 □ Vfinal

100
80
60
40
20
0

freq. verbs in test   correct agrammatic speakers

□ V2 □ Vfinal

overall   definite object   test1   test2

□ adjunct-object □ object-adjunct

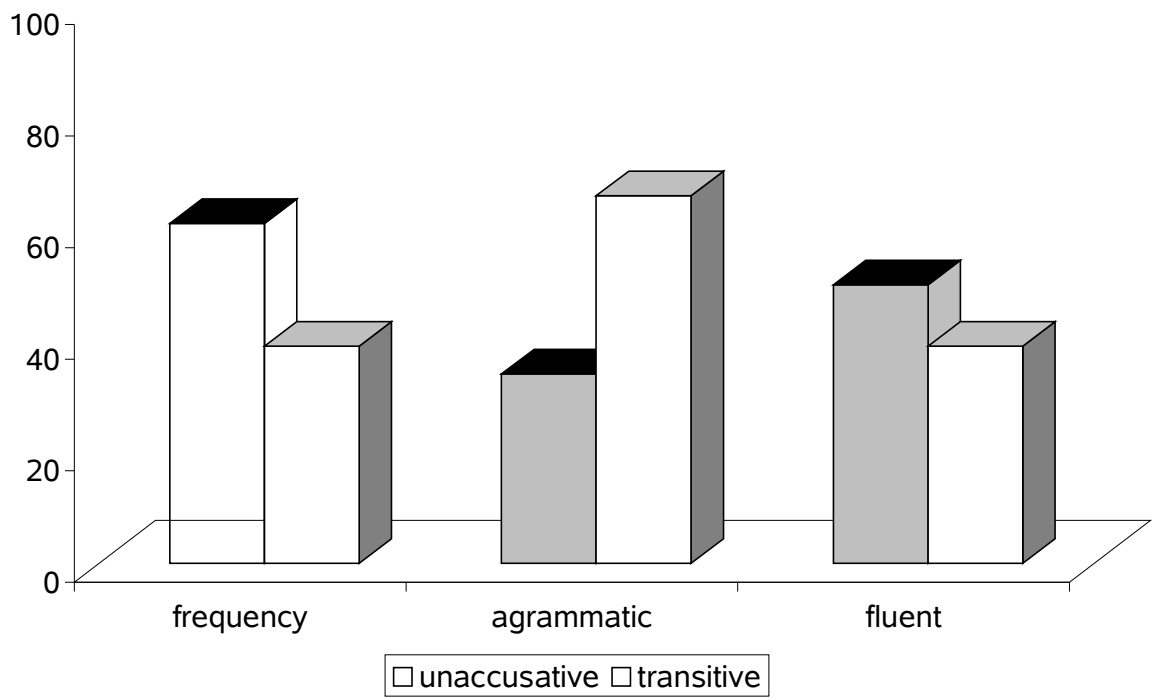|	frequency	agrammatic	fluent
unaccusative			
transitive			

Table 1a: Absolute and relative frequencies of the verbs in Verb Final (=base) and Verb Second position in the Corpus of Spoken Dutch. For clarification of the categories: see text.

Comparison	LEX	FIN	OBJ	Verb Final		Verb Second	
				count	%	count	%
(1)	+	–	+	21,796	52.1	20,073	47.9
(2)	–	+	–	18,446	20.4	71,755	79.6
(3)	+	+	–	13,155	19.3	54,837	80.7
(4)	+	+	+	4,865	21.5	17,717	78.5

Table 1b: Absolute and relative frequencies of the verbs in Verb Final (=base) and Verb Second position in newspaper data (*Algemeen Dagblad*, January, 1999).

Comparison	LEX	FIN	OBJ	Verb Final		Verb Second	
				count	%	count	%
(1)	+	–	+	23,327	59.7	15,749	40.3
(2)	–	+	–	24,056	25.9	69,066	74.1
(3)	+	+	–	15,706	25.3	46,628	74.7
(4)	+	+	+	5,954	27.7	15,669	72.3