Animacy and Canonical Word order – Evidence from Human Processing of Anaphora

Espen A. Larsen
LLE
University of Bergen
Bergen, Norway
espen.aleksander@gmail.com

Christer Johansson
Computational Linguistics
University of Bergen
Bergen, Norway
christer.johansson@uib.no

1 Introduction

Previous studies have shown that canonical word order is processed faster than non-canonical word order (Menn, 2000; Kaiser and Trueswell, 2004, inter al.). Both SVO and OVS word order are possible in Norwegian, but the SVO word order can be considered the canonical word order. We confirmed this for Norwegian in an early experiment (Larsen, 2005) where faster reaction times were recorded for SVO, and the SVO word order was also preferred for ambiguous sentences.

The initial results support Lyn Frazier's Garden Path Theory (Frazier, 1994), which states that only one syntactic possibility is tried at a time, and the first option is to try the simplest structure first, in accordance with the Minimal Attachment Principle. Thus, the choice of the syntactically simplest structure should not be influenced by its semantic content.

In speech, prosody aids functional role assignment, and correlates with new/given information (Horne and Johansson, 1993). In text we need to rely on other means of disambiguation, one being tracking the status of new or given information (ibid). Other factors for understanding referentiality have been discussed and proposed (Foss and Ross, 1983; Jarvella and Engelkamp, 1983; Garrod and Sanford, 1983; Garnham, 1984; Garrod, 1994, inter al.). The experiments we present here concern processing of sentences, and interaction between word order and semantic distance. Since word order is inherently ambiguous in Norwegian we use a context that clarifies the intended functional roles of the next sentence.

The experiment we present here considers semantic content as well as the word order. According to our alternative hypothesis, we expect to see effects of animacy on reaction time and accuracy. The null hypothesis is that there is no such effect of animacy.

We compared matched sentences containing exactly the same words, to control for lexical frequencies. We simultaneously controlled for differences between positive and negative responses. In order to do this we used a clarifying question, which disambiguated the word order of the following context sentence, after which we presented a probe sentence that was either a natural continuation of the previous context, or an unnatural continuation. This is presented in the experimental set up in table 3.

2 Design

The experiment is designed to deal with unambiguous word order alterations. The key to this is a clarification question which specifies the subject of the following answer. The answer is stated either as a SVO or an OVS sentence. This is followed by a sentence which the subject is asked to judge natural continuation or not. We tested the effect of word order by taking the difference between the b-conditions (OVS) and the a-conditions (SVO) (cf. table 1). Furthermore, we controlled for positive and negative answers as well as the verb in the continuation sentence. We compare the following differences in reaction time:

(1b-1a) [neg. *den plystret* / it whistled],
(2b-2a) [pos. *hun plystret* / she whistled],
(3b-3a) [pos. *den bjeffet* / it barked] and
(4b-4a) [neg. *hun bjeffet* / she barked].

Groups 1 to 4 defined the logical possibilities of two factors: expected answer (yes/no) and if the previous object was in subject position or not in the continuation. We constructed corresponding stimulus for three different levels of animacy (inanimate, animal, and human). The subject of the context sentence was always human. Positive and negative responses may differ for these specific comparisons. Note that all lexical content is exactly the same in each comparison. Hence, the design cancels out lexical

Johansson, C. (Ed.)
Proceedings of the Second Workshop on Anaphora Resolution (2008)

effects, leaving us with a pure effect of word order of the context sentence. Repetition effects will be canceled out, as each subject is presented with her own random order of repetitions.

2.1 Related experiment

2.1.1 Weskott 2003

Weskott (2003) considers German SO (Subject-Object) and OS (Object-Subject) structures, in unambiguous contexts. Weskott used critical sentences that were made unambiguous by German case marking. Previous research indicated that context does not play a strong enough role to compete with preferred syntactic structure in sentence processing. Weskotts experiment investigates inferable vs. non-inferable (possible/not possible to infer S and O from context) SO and OS structures. Weskotts experimental structure was quite similar to the one we chose, and likewise confirmed an effect of inferability from context, i.e. discourse processing.

His results show that non-inferable OS structures cause most processing problems; higher reading time and lower accuracy. As we can see, his experimental design was quite similar to the one we chose (compare table 1 and 2), but differing in the fact that it was self-paced (i.e. the participants pressed a button when they wanted to read the next word), while our design presented a whole sentence at a time.

2.1.2 The Information Load Hypothesis

A full NP anaphor (such as *bird*) is read faster if it has a more typical antecedent, e.g. *robin* compared to *goose* (Garrod and Sanford, 1977). This typicality effect can be reversed if the antecedent was presented as a clefted constituent (Almor, 1999; Wind Cowles and Garnham, 2005). Almor (1999) proposed the Informational Load Hypothesis (ILH) to account for both similarity effects, and the interaction with syntactic construction and focus. High activation of a focussed item may compensate for a lower level of conceptual similarity.

> "Because the informativeness of a referring expression is dependent on what is already known, the informational load of an anaphor is defined with respect to its antecedent and is determined by the difference between the semantic representation of the anaphor and the semantic representation of the antecedent." (Almor, 1999, p.751)

Almor (ibid) discusses a measure called C-difference "as a formal link between the informational load of the anaphor-antecedent pair and the semantic distance between the anaphor and the antecedent." A C-difference can be positive (when the anaphor is less general) or negative (when more general). The C-difference is set up as a monotonous function of semantic distance, which then gets its sign from the more general: antecedent (+) or anaphor (−).

Almor defines Informational Load as a relation with C-difference such that if the C-difference between $anaphor_1$ and $antecedent_1$ is larger than between $anaphor_2$ and $antecedent_2$, so is the informational load. No explicit formula is provided for calculating a value for informational load, and it is generally not possible to give such an explicit calculation since we cannot access the semantic distances. However, if we keep either the anaphor or the antecedent constant we may determine which pair has the highest informational load.

> "In sum, the ILH claims that the ease of processing NP anaphors can be described by the interaction of three factors: discourse focus, the amount of new information added by the anaphor, and the informational load of the anaphor-antecedent pair." (Almor, 1999, p.753)

In our own experiment we are testing word order effects (SVO or OVS) at the sentences that introduce the antecedent (e.g. *the woman walked the dog*). The word order gives a syntactic focus on either the object or the subject, i.e. the object could have been raised to a topicalized first position. The anaphor either relates to the previous object or subject, and the verb of the anaphor gives a semantic focus on either the previous object or subject. The pronoun should be enough to select the correct antecedent, but the various cues interact and we can show effects of word order, and weaker effects for semantic focus. Thus we can see a relation between our experimental set up, and the ILH. One difference, however, is that we are in effect comparing effects related to the distance between two possible *antecedents* keeping the anaphor constant for each paired comparison and always keeping one antecedent in the *human* class.

Table 1: The experimental stimuli. *Example from human subject and animal object.*

		Clarifying Question	**Context (SVO/OSV)**	**Continuation**	**Expected Answer**
		What (obj)... woman (subj)?	*She walked the dog.*	whistled/barked	
1a	**SVO**	Hva luftet kvinnen?	Kvinnen luftet hunden.	Den plystret.	NO
2a	**SVO**			Hun plystret.	YES
3a	**SVO**			Den bjeffet.	YES
4a	**SVO**			Hun bjeffet.	NO
1b	**OVS**	Hva luftet kvinnen?	Hunden luftet kvinnen.	Den plystret.	NO
2b	**OVS**			Hun plystret.	YES
3b	**OVS**			Den bjeffet.	YES
4b	**OVS**			Hun bjeffet.	NO

Table 2: Weskott's design (Weskott, 2003, p.65)

(1) **Lead-in sentence:**
Peter freute sich auf seine Mittagspause.
Peter was looking forward to his lunch break.

(2) **Context sentence:**
Er ging in die kleine Pizzeria in die Innenstadt, wo allerdings eine feindselige Stimmung herrschte.
He went to the small pizzeria in the inner city, where however a hostile atmosphere reigned

(3) **Critical sentence:**
(a) Der Kellner beleidigte den Koch ziemlich heftig.
The waiter (NOM) insulted the cook (ACC) pretty intensely
(b) Den Kellner beleidigte der Koch ziemlich heftig.
The waiter (ACC) was insulted by the cook (NOM) pretty intensely
(c) Der Metzger beleidigte den Koch ziemlich heftig.
The butcher (NOM) insulted the cook (ACC) pretty intensely
(d) Den Metzger beleidigte der Koch ziemlich heftig.
The butcher (ACC) was insulted by the cook (NOM) pretty intensely

3 Procedure

Our experiment presented full sentences. The clarifying question and the context sentence were presented for a 1000 milliseconds. The continuation sentences were presented for 1700 milliseconds, but the presentation ended when the subject gave his or her response. The subjects were instructed to consider if the continuation sentence seemed natural or unnatural given the previous context, pressing no for unnatural or yes for natural. This was the decision task for each sentence sequence (1a, 2a, 3a, etc.)

The sequence groups consisted of sentences containing the following three participant combinations:

(1) Human – Inanimate
(2) Human – Animate
(3) Human – Human

A context sentence would always contain a human participant. The other NP would randomly vary between being human, animate or inanimate, with equal presentation frequency. The presentation order was randomized by E-Prime, to avoid that subjects could figure out a presentation pattern. For example, sequence 1a from one sequence group would be followed by sequence 4b from another sequence group, with a different content and different semantic relations. We included a training list containing 24 samples, and the trial list contained 120 trials (i.e. sequences).

Each test subject sat in a chair in front of a computer screen in a soundproof room. First, the experiment instructions were given orally. When the subject started the program, instructions were repeated on the screen, and the subject confirmed having read instructions before starting the actual test. This gives an opportunity to ask questions, if anything was unclear, before starting the experiment.

The program presents one whole sentence at a time. One sequence consists of three stimuli sentences. The clarify-sentence remains on the screen for 1000 ms (milliseconds, i.e. 1 second), then a blank screen for 50ms, followed by the related-sentence, which is presented for another 1000 ms, a blank screen for 50ms, and finally the related-sentence is presented until a decision is made or time out at 1700ms. A preliminary pilot test had shown that the test subjects could process the information and make a decision in less than this time. If the subject hesitates too long, there will be a time-out and thus no recorded reaction time for that sequence, however most subjects had no time-outs. The subject pressed a button for either natural or unnatural continuation, given the sentences of each sequence. Responses were given through an E-Prime Deluxe Response Box for reliable measurements.

The experiment took between 17 and 18 minutes on average for each subject. To help the subjects to keep their focus on the task, we had inserted a break in the program half-way through the trial set. After the break, a few training sequences were shown to get the subject into the task again.

The statistical analysis compared sentences differing only in the word order of the context sentence (SVO or OVS). Statistical differences would then be due to this word order difference, everything else being the same (paired comparisons 1a and 1b, 2a and 2b, etc.) (see Table 1). The design gives us a very good opportunity to study pure word order differences.

For example comparing reaction times from 1a and 1b in table 1 makes it obvious that everything, from the lexical elements to the expected decision, is the same except for the word order of the context sentence. We used as factors the expected decision (and only included responses that agreed with the expected decision) and the focus of the verb in the continuation (which could have as it subject either the previous subject or the previous object).

Participants who failed to get a high enough performance were excluded. Some subjects used almost exclusively one answer, and some had too many time outs. We excluded all non-expected answers, after which we performed outlier detection, and excluded all measurements that were outside of 2.5 standard deviations. We did the same for excluding extreme differences between paired measurements. Without the outlier detection the results would be slightly stronger, but we wanted to make sure that the results were not only due to a few extreme observations. The remaining data contained 1079 paired data points, with recorded reaction time differences for 27 subjects. In our experiment, the currently active discourse representation is limited, and randomly varies between SVO and OVS structure. The ordering of stimuli is different between participants, in order to cancel out effects of presentation order. The presentation was programmed so that half of the time the OVS version was presented first, and half of the time the SVO version came first.

4 Results

The experiment showed (cf. figure 1) a significant effect of animacy ($F(2, 1067) = 4.56$, $p < 0.02$) using 1079 paired data points obtained from 27 subjects (university students, and a few staff). Processing is significantly faster for OVS-context when the test word is animate, and human is significantly slower with this presentation. However, inanimate words are not significantly affected. This supports a rejection of our null hypothesis, and we may conclude that there is an effect of animacy for the preferred word order of contextual information. The animacy effect might be an effect of searching for an antecedent (i.e. after parsing the context), or it might be a direct effect of semantics on word order preferences (i.e. was a reparse always necessary for OVS?). Both interpretations give rise to interesting questions.

Interaction between word order and the verb of the test sentence, which may focus on either the previous subject or the previous object, approaches significance: F(2,1067)=2.69, p<0.07.

Accuracy differs between the different categories (cf. table 3). Each category has an equal number of presentation, so we would expect the same number of mistakes in all categories. We have chosen to analyze the differences between OVS and SVO for each animacy group. The Human category has more, the Inanimate cate-

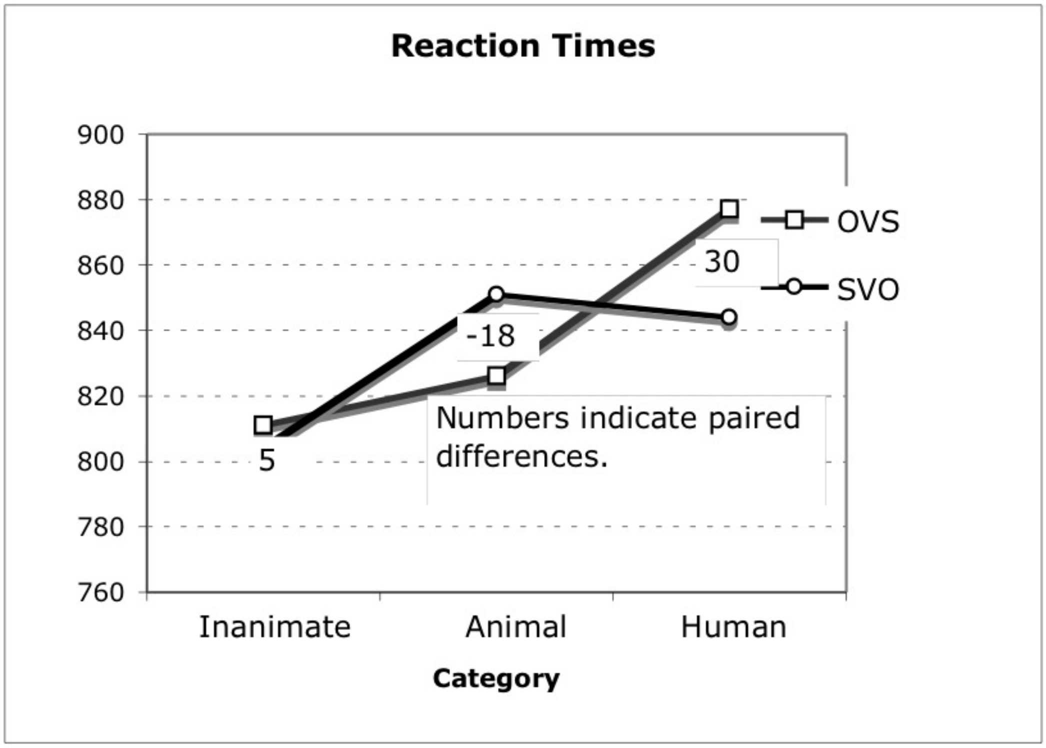

Figure 1: Effect of Animacy

Table 3: Accuracy in the different categories.

	Inanimate	**Animal**	**Human**
OVS	418	468	392
SVO	437	492	444
Difference	**19**	**24**	**52**
Expected	**31.7**	**31.7**	**31.7**
effect	**-12.7**	**-7.7**	**20.3**
χ^2	5.09	1.87	13.00

gory less, incorrect answers than an equal distribution would predict. There are different accuracies in the different groups; the overall differences are highly significant ($\chi^2 = 19.96$, $df = 2$, $p < 0.0001$), and most of the effect shows up in the human category, which is the category with the least cognitive/semantic distance between the antecedent candidates.

The results on accuracy can be illustrated in a Cohen-Friendly plot (see figure 2). The width of each bar is proportional to the squared root of the expected frequency (proportional to association strength), and the height is proportional to the χ^2 contribution to each cell (i.e. significance). The area of each bar is proportional to the difference of observed and expected frequencies.

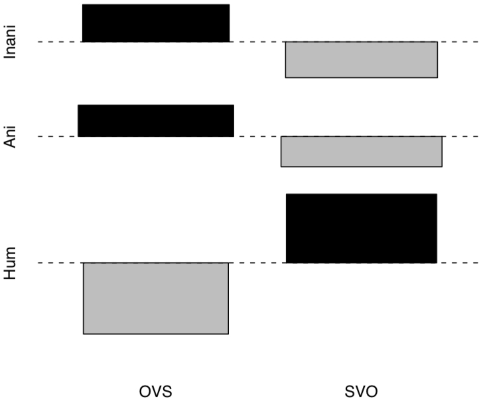

Figure 2: Cohen-Friendly plot of the effect of animacy on accuracy

5 Conclusion

We found a significant effect for the effect of animacy on response time and accuracy in the processing of SVO and OVS sentences. This could be an indication that semantics play a part at an early stage of sentence processing, and it is a definite argument against the Modularity Theory, since interaction between semantic information and word order information can be observed in reaction times. The subjects made significantly more mistakes in the Human-Human category, compared to Human-Inanimate and

Human-Animate.

The fact that animacy gives a significant effect for both response time and accuracy is an argument against an interactive processing mechanism. On the other hand, Frazier's theory states that SVO should always be the preferred word order. In this experiment, sentences with a topicalised animate object were processed faster than sentences with an animate object in the NP2 position.

Our results could be evidence for a limited interactive system. The Human-Human sequences put an extra burden on the parser in the shape of too much different information, and this leads to a higher processing time and a lower accuracy in the critical decision process. If we compare the human parser to a computer we see the same result; the more parallel processes we have going on, the slower they are performed. The design of our experiment guarantees that the effects we found can only be related to the word order difference in the context sentences (SVO/OVS). It is a pure word order effect. All other possible factors are controlled for through the design.

The results can, to some extent, be explained on the basis of Antecedent Identifiability (Garrod, 1994). Garrod says that one can establish a continuum in terms of the degree to which anaphora identify their discourse antecedents. The more explicit the expression is, the more rapidly the antecedent-anaphora relationship is solved.

Garrod's theory states that it is the contextual presuppositions of the anaphora that actively drives the resolution process. On this basis, an efficient sentence-resolution system should attempt to incorporate the interpretation of the sentence and its elements directly into the currently active discourse representation.

> ".....in most contexts, a pronoun by itself could map onto many different textual antecedents. Demonstrative and other definite NPs have more semantic content, but they are also commonly used to identify antecedents that do not uniquely match this content (Garnham, 1984; Garrod and Sanford, 1977). However, when used anaphorically, names usually serve as unique identifiers. Thus, the different devices range along an explicitness continuum in terms of the degree to which they uniquely identify their discourse antecedents. As is seen later, this difference has been used to account for apparent differences in the time course of resolution of the different types of anaphora, where it has been suggested that the more explicit the expression the more rapidly it is resolved (Gernsbacher, 1989)." (Garrod, 1994, pp. 345–346)"

Why does it affect the response accuracy if the second noun phrase is inanimate, animate or human? The semantic relations are already given in the clarification question. An explanation to this must probably be related to working memory. The Hum-Hum sequences are both grammatical and unambiguous (given clarification), but they still generate a higher reading time and a lower accuracy. An explanation could be that they need more cognitive processing resources than what are available. This structure gives rise to reprocessing and places a heavier burden on working memory an argument for Garden Path and Modularity of Mind (Crocker and Keller, 2005).

Our results show that semantic distance affects anaphor resolution, in line with the Information Load Hypothesis. The Human-Inanimate category was processed faster than the other two categories, and with a higher accuracy, because the antecedent candidates were clearly distant in the cognitive hierarchy. Integrating new information with given information may also be easier the more separate the semantic roles are.

The fact that there is a difference between Hum-Ani and Hum-Hum sentences is partly an argument for and partly an argument against Frazier's Garden Path Theory. The fact that the effect of animacy is significant, supports Fraziers theory. But, the fact that OVS sentences are processed faster for the Human-Animate category (see figure 1) is an argument against the same theory, which states that the parser always tries the simplest (most frequent) structure first.

Acknowledgement

This work is based on the first author's MPhil thesis. We thank professor Robert Jarvella for good suggestions for further research. We also thank Gard Buen Jenset for suggesting relevant statistical illustrations, and for his help with R.

References

A. Almor. 1999. Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, 106(4):748–765.

M.W. Crocker and F. Keller. 2005. Probabilistic grammars as models of gradience in language processing. In G. Fanselow, C. Fery, R. Vogel, and M. Schlesewsky, editors, *Gradience in Grammar: Generative Perspectives.* Oxford University Press, Oxford, UK.

D. J. Foss and J. R. Ross. 1983. Great expectations: Context effects during sentence processing. In G.B. Flores D'Arcais and R.J. Jarvella, editors, *The Process of Language Understanding*, pages 169–191. John Wiley & Sons.

L. Frazier. 1994. Introduction. In C. Clifton, L. Frazier, and K. Rayner, editors, *Perspectives on Sentence Processing*, pages 1–12. Lawrence Erlbaum Assoc., Hillsdale, NJ.

A. Garnham. 1984. Effects of specificity on the interpretation of anaphoric noun-phrases. *Quarterly Journal of Experimental Psychology*, 36A:1–12.

S. Garrod and A.J. Sanford. 1977. Interpreting anaphoric relations: The integration of semantic information while reading. *Journal of Verbal Learning and Verbal Behavior*, 16:77–90.

S. Garrod and A. Sanford. 1983. Topic dependent effects in language processing. In G.B. Flores D'Arcais and R.J. Jarvella, editors, *The Process of Language Understanding*, pages 271–296. John Wiley & Sons.

S. Garrod. 1994. Resolving pronouns and other anaphoric devices: The case for diversity in discourse processing. In C. Clifton, L. Frazier, and K. Rayner, editors, *Perspectives on Sentence Processing*, pages 339–357. Lawrence Erlbaum Assoc.,, Hillsdale, NJ.

M. A. Gernsbacher. 1989. Mechanisms that improve referential access. *Cognition*, 32:99–156.

M. Horne and C. Johansson. 1993. Computational tracking of new vs. given information: implications for synthesis of intonation. In B. Granström and L. Nord, editors, *Proc. Nordic Prosody VI—papers from a symposium*, pages 85–97. Almquist & Wiksell International, Stockholm, Sweden.

R. J. Jarvella and J. Engelkamp. 1983. Pragmatic influences in producing and perceiving language: A critical and historical perspective. In G.B. Flores D'Arcais and R.J. Jarvella, editors, *The Process of Language Understanding*, pages 225–270. John Wiley & Sons.

E. Kaiser and J.C. Trueswell. 2004. The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94(2):113–147.

E. Larsen. 2005. V2 in Norwegian – Consequences for Sentence Processing. (ms. University of Bergen, Norway).

L. Menn. 2000. It's time to face a simple question: What makes canonical form simple? *Brain and Language*, 71:157–159.

T. Weskott. 2003. *Information Structure as a Processing Guide: The Left Periphery of German Verb-Second Sentences and Its Interpretation in Context.* Ph.D. thesis, University of Leipzig, Leipzig, Germany.

H. Wind Cowles and A. Garnham. 2005. Antecedent focus and conceptual distance effects in category noun-phrase anaphora. *Language and Cognitive Processes*, 20(6):725–750.