Sentence processing theories and the position of complex constituents in Dutch texts*

Frank Jansen
Utrecht University/UiL-OTS

1. Introduction

A sentence like (1) is far from abnormal in newspaper Dutch:

(1) (Kop) Olympische missie delicaat probleem voor Ritsma
In de herfst van zijn loopbaan, waarin hij het vizier meer dan ooit gericht heeft op die ene medaille die hem nog ontbreekt, weet Rintje Ritsma zich voor een delicaat olympisch vraagstuk geplaatst.[1]
(Headline) Olympic mission delicate problem for Ritsma
In the autumn of his career, in which he has caught sight of the only medal that is lacking to him more than ever, knows Rintje Ritsma himself confronted with a delicate Olympic problem.

A striking aspect of (1) is the fact that the long and complex adverbial phrase *In … ontbreekt* is in first position, while it could easily have been placed at the end of the sentence (1a), or at least somewhere in the middle of it (1b):

(1) a. Rintje Ritsma weet zich voor een delicaat olympisch vraagstuk geplaatst in de herfst van zijn loopbaan, waarin hij het vizier meer dan ooit gericht heeft op die ene medaille die hem nog ontbreekt.
b. Rintje Ritsma weet zich in de herfst van zijn loopbaan, waarin hij het vizier meer dan ooit gericht heeft op die ene medaille die hem nog ontbreekt, voor een delicaat olympisch vraagstuk geplaatst.

According to Behaghels *Gesetz der wachsenden Glieder* (Behaghel 1909) we would expect the variants (1a) and especially (1b) to be stylistically superior to (1), as the adverbial phrase is by far the longest constituent of the sentence. However, this expectation would not be in accordance with our intuitions. While (1a) is stylistically flawless, even better than (1), (1b) is intuitively inferior to both (1) and (1a). What seems to trouble the reader of (1b) most, is the problem how to succeed after

Linguistics in the Netherlands 2003, **83–92.**
ISSN 0929–7332 / E-ISSN 1569–9919

reading *ontbreekt*. It seems difficult for him 'to dive up' from this relative clause in a relative clause in an adverbial phrased, in order to interpret the next constituent as a part of the main clause.

The correct generalization seems to be that complex constituents avoid the middle position and prefer to be placed outside the verbal bracket, which consist of the tensed verb in second position (*weet*) and the participle at the end (*geplaatst*). Considering the remaining two options at the borders of the sentence, a complex constituent on the last position makes a better impression than on the first position.

Do recent sentence-processing theories predict this order of acceptability? This question is the topic of this contribution. I will analyse sentences like (1) with the help of the processing theories proposed by Hawkins (1994; 1998) (see Section 3) and by Gibson (1998, 2000) (Section 4). But before it comes to that, some data on the position of complex phrases in Dutch texts are presented to give a more substantial underpinning of the evaluations above (Section 2).

2. The position of complex phrases in Dutch

Avoid brackets constructions! That is one of the oldest and most common text advices in Dutch normative linguistics (see Renkema 1981: 38–39 and Maureau (1983) for analyses of the older literature). The two verbal elements, viz. the tensed verb in second position and the other verbal elements in (pre-)final position, are considered as a pairs of brackets, which are difficult to associate with each other when there are, more (or longer, or more complex) constituents in between. The advice is to get rid of the bracket construction by placing the intervening elements outside the bracket, especially towards the final position.

The first position is a less obvious candidate for these constituents, because text advisers do not like sentences beginning with long or complex constituents other than subjects either (see for example Renkema 1995). They consider sentences with a so-called *lange aanloop* 'long introduction' to be complex as well. However, the advice to avoid sentences with long introductions is less firmly established in the tradition of normative grammar than the bracket advice. For example it is not included in the surveys by Renkema (1981) and Maureau (1983). So we may conclude that (1a) is preferred to (1) and (1) is preferred to (1b) in normative grammar.

Besides this evidence from normative grammar, there is also evidence from text frequency. I counted simple and complex constituents occupying the three positions (first, middle and last) in five newspaper texts.[2] Complex constituents were defined in two ways:

a. the constituent had at least one relative clause (like *die hem nog ontbreekt* in (1)),
b. the constituent had at least one prepositional phrase in post nominal position (like *van zijn loopbaan* in (1)).

Simple constituents contained neither a relative clause nor a post nominal PP. Some constituents are excluded from occupying the last position by grammatical constraints, such as pronouns and syntactic subjects. Those constituents were filtered from the corpus.

Firstly we would like to know whether complex constituents avoid the middle position. See Table 1 for an answer.

Table 1. The frequency of complex and simple constituents in the middle position or at the borders of main clauses

	middle position	
	yes	no
Complex constituents	26 (22%)	91 (78%)
Simple constituents	149 (65%)	81 (35%)

Table 1 demonstrates that complex phrases dislike a middle position, and prefer a position outside the verbal bracket instead, when compared with simple constituents. The difference in frequency between simple and complex constituents statistically significant ($\chi^2 = 56.19$; $p < .0001$).

Secondly, we would like to know whether complex constituents occupy more frequently the last position than the first position. See Table 2 for an answer.

Table 2. The position of complex and simple constituents at the borders of main clauses

	borderline position	
	first	last
Complex constituents	28 (31%)	63 (69%)
Simple constituents	43 (53%)	38 (47%)

The results presented in Table 2 demonstrate that complex constituents prefer the last position above the first position when compared with simple constituents ($\chi^2 = 8.81$, $p < .003$).

My conclusion is that complex constituents prefer the last position and eschew the middle position, while the first position is somewhere in between.

3. Hawkins' Early Immediate Constituents Theory

Hawkins' reasoning behind his Early Immediate Constituents theory (1994) is relatively straightforward. Successful parsing depends on the first moment when the human parser is able to know what the structure of the sentence will be: the earlier that moment, the simpler the parsing process. In order to oversee the sentence structure, the parser has to wait until he has reached the last constituent of the sentence. However, this does not mean that he has to wait until he has reached the last word of the last constituent. The decisive moment comes earlier, viz. when the parser has got enough information to know what the structure of the last constituent will be. As soon as he has reached this word the parser will build up the structure of the last constituent *and* of the entire sentence.

An example will suffice to understand how Hawkins' theory works. In (1) and (1b) the parser has to wait until the last constituent, which is also the last word of the sentence (*geplaatst*), before he can process the sentence. In (1a) however, he has only to wait for *in*, the first word of the last constituent. From *in* he infers that what follows has to be an adverbial phrase and nothing else, and he can safely assume that he has already processed the verbs and their arguments. It is easy to see that (1a) is less complex than (1) and (1b) according to Hawkins' procedure, *in* being the tenth word and *geplaatst* the 31st word. It goes without saying that Hawkins' theory favours word orders with the most complex constituent on the last position.

Hawkins has also devised a procedure to predict the relative order of complexity of non-optimal variants. This time, his way of computation is more complex and less straightforward. To compute the ranking order of two non-optimal variants, Hawkins proposes a procedure comprising the followings steps:

– assign the constituents of the sentence a ranking order from left to right,
– assign the words of the sentence a ranking order from left to right,
– divide the ranking order of each constituent by the ranking order of its last word,
– compute the mean of these divisions for all constituents.

The sentence variant with the highest mean is the best of the non-optimal variants.

When we apply this procedure to the non-optimal variants (1) and (1b), we get:

(1) $1/23 + 2/24 + 3/26 + 4/27 + 5/32 + 6/33 = 0.73/6 = 0.12$
b. $1/2 + 2/3 + 3/4 + 4/27 + 5/32 + 6/33 = 2.41/6 = 0.4$

Hawkins' procedure results in the prediction that (1b) is more close to the optimal variant than (1), while our intuitions and the evidence presented in Section 2 arrived at the opposite order.[3]

It is highly improbable that this result can be attributed to the particulars of (1), as the procedure favours constituent orders with constituents increasing in complexity

from left to right. In other words, Hawkins' procedure is a kind of formalization of the Gesetz der wachsenden Glieder, while a peculiarity of Dutch word order is that complex constituents are better in initial position than in middle position.

We have to conclude therefore, that Hawkins' (1994) theory is fine for selecting the optimal variant, but systematically wrong for the variant with a complex constituent on the first position.[4]

4. Gibson's Syntactic Prediction Locality Theory

Gibson's Syntactic Prediction Locality theory (Gibson 1998) is more complex than Hawkins' theory. He assumes that the human parser comprises two activities:

– the integration of every next word in the sentence into the structure built up so far,
– the storage in working memory of all elements that are needed to complement the elements processed so far for a meaningful grammatical sentence.

Sentence complexity is a function of the costs of those two activities of the human parser. In this paper I will focus on the integration costs, because my computation of the memory costs resulted in rather unsatisfactory outcomes. When I computed the memory costs in the SPLT framework of Gibson (1998), the result turned out to be a copy of the integration costs. Then I applied the way of computation memory costs that Gibson (2000) proposes in Local Dependency Theory, a revised version of SPLT, which resulted in outcomes that were very difficult to interpret.

4.1 Integration

An important aspect of Gibson's integration theory is that it focuses on the verbal elements and their arguments. Integration means:

– attaching the new element to the correct head,
– assigning correct syntactic and semantic roles to the argument NP's.

The costs of these integration activities depend on the distance between the new element and its head. Gibson measures this distance in new discourse referents (nouns and verbs) that have to be processed after the head, including the new element. Every new discourse referent costs one Integration Unit (IU). The more IU's are needed at a certain point in the parsing of the sentence, the more complex the sentence is at that position.

How Gibsons theory works is demonstrated by applying it to an example: (2), a simplified version of (1). The numbers are the IU's needed to integrate this constituent:

(2)

In	de	herfst	van	zijn	loopbaan	ziet	Rintje	een	dilemma
0	0	0	0	0	0	0	IU(1)	0	IU(2)

Every element of the PP *In … loopbaan* is processed by attaching it to the previous element, so there are no integration costs involved here. This changes directly after the tensed verb *ziet* 'sees'. As *to see* is a transitive verb, the parser knows that a subject and an object will follow. When *Rintje* is integrated, it is attached to *ziet*, at the cost of one IU of one element (notated one IU(1)). This is because the parser has to process one constituent, viz. *Rintje,* to integrate *ziet.* Parsing the next constituent, the NP *een dilemma,* means also integrating it with *ziet.* This time the integration cost is one IU(2), because the parser has to pass *Rintje* and *een dilemma.*

The integration costs of the variants of (2) are:

(2) a.

Rintje	ziet	een	dilemma	in	de	herfst	van	zijn	loopbaan
0	IU(1)	0	IU(1)	0	0	0	0	0	0

b.

Rintje	ziet	in	de	herfst	van	zijn	loopbaan	een	dilemma
0	IU(1)	0	0	0	0	0	0	0	IU(3)

The optimal variant for integration is (2a): it costs only two IU(1)' s: one IU to integrate the subject to *ziet* and one IU to integrate the object to *ziet.* Variant (2b) is more complex: the parser needs one IU(1) to integrate the subject to the verb, and one IU(3) to integrate the object with *ziet.* After *ziet*¸ the parser has to process *herfst, loopbaan* and *dilemma* in order to be able to interpret *een dilemma* successfully as the object of *ziet.*

A comparison of the integration costs of the variants reveals that their ranking order of complexity is: (2a) < (2) < (2b), a result that is in accordance with the evidence presented in Section 2.

4.2 Bracket structures

What does Gibson's SPLT predict in the case of the bracket construction exemplified by (1b). The answer is given in the following analyses of another simplified variant of (1).

(3)

In	de	herfst	van	zijn	loopbaan	moet	Rintje	een	dilemma	oplossen
0	0	0	0	0	0	0	0	0	0	IU(3)+ IU(2)+ IU(1)

The parser has to wait until the main verb *oplossen* before he is able to integrate the auxiliary *moet* (which is three referents away, hence IU(3)), the subject *Rintje* (IU(2)) and the object *een dilemma* (IU(1)).

(3) a.

Rintje	moet	een	dilemma	oplossen	in	de	herfst	van	zijn	loopbaan
0	0	0	0	IU(3)+ IU(2)+ IU(1)	0	0	0	0	0	0

b.

Rintje	moet	in	de	herfst	van	zijn	loopbaan	een	dilemma	oplossen
0	0	0	0	0	0	0	0	0	0	IU(5)+ IU(4)+ IU(1)

The ranking order that Gibson's computation of integration costs predicts, is: (3a) and (3) < (3b). We may conclude that the high complexity of the variant with the large bracket construction is correctly predicted. On the other hand, the procedure fails to differentiate between the two variants with the complex constituent at first or last position.

So, both theories offer only a partial explanation of the complexity of the Dutch constituent order: Hawkins explains why complex constituents prefer the sentence final position. Gibson explains why those constituents avoid the middle position. It seems to me rather unattractive from a conceptual point of view to explain one phenomenon with two different theoretical frameworks. Therefore I will try to adjust Gibson's theory somewhat.

4.3 A tentative solution

What went wrong with the computation of the integration costs in the previous section? My answer is that the integration totally depends on the main verb. I think it is not realistic to assume that the parser has to wait for the main verb at the end of the clause and refrains from any integration activities while processing the tensed verb.

In Dutch main clauses the position of the tensed verb is fixed in second position. The position of the subject is fixed as well: either directly before the tensed

verb or directly behind it.[5] In other words, when the parser realizes that the first constituent cannot be the subject, he may safely assume that the constituent after the tensed verb is the subject. I propose that at that position the integration of the subject takes place already.

(4)

In	de	herfst	van	zijn	loopbaan	moet	Rintje	een	dilemma	oplossen
0	0	0	0	0	0	0	IU(1)	0	0	IU(3)+ IU(1)

a.

Rintje	moet	een	dilemma	oplossen	in	de	herfst	van	zijn	loopbaan
0	IU(1)	0	0	IU(2)+ IU(1)	0	0	0	0	0	0

b.

Rintje	moet	in	de	herfst	van	zijn	loopbaan	een	dilemma	oplossen
0	IU(1)	0	0	0	0	0	0	0	0	IU(4)+ IU(1)

This approach leads to an ranking order of: (4a) < (4) < (4b), which is in accordance with the evidence presented in Section 2.

5. Conclusions

Let us start the conclusions by looking back upon the original fragment (1) and its variants. When we compute the integration costs it is the participle *geplaatst* where we find significant differences in integration costs for the auxiliary: They vary from one IU(2) in (1a), via one IU(3) in (1), to a hardly acceptable IU(7) in (1b). This result seems to me a good reflection of the intuitive evaluations of the variants.

In other words, the advantage of the proposed revision of the computation of integration costs is that it is in accordance with intuitive ideas and text frequencies. Another advantage is that one does not have to be driven back on two theories: Hawkins' EIC for explaining the superiority of sentence variants with complex constituents at the end, and Gibson's SPLT for explaining why variants with a complex constituent in initial position are superior to variants with that kind of constituents in middle position.

However, the revision has at least one disadvantage as well: The concept of what integration means is somewhat blurred. In Gibson's SPLT and DLT successful integration means attribution of the all syntactic *and* semantic functions to all arguments of the verb. This is attractive because both (and especially the semantic functions) are needed as a basis for the construction of a mental model of the text.

A consequence of the proposed revision is that syntactic integration can take place at an earlier moment than semantic integration, what makes the parsing model less simple. So it remains to be seen whether there are no alternative ways for accounting for the evidence presented in Section 2. Furthermore we need data from reading experiments to see whether the hypothesized time-consuming integration activities do take place when the tensed verb is read. Dutch is very apt for this kind of experiments, as it has tensed verbs in second position (in main clauses) and in last position (in dependent clauses). Finally the alternative ways of computing integration costs have to be applied to other languages with bracket structures.

Notes

* I thank dr. J.A. Hawkins, dr. H. Pander Maat, drs. R. Wijnands and an anonymous LIN reviewer for their helpful comments on an earlier version.

1. (1) is a slightly altered version of the original (i):

(i) In de herfst van zijn loopbaan, met het vizier meer dan ooit gericht op die ene medaille die hem nog ontbreekt, weet Rintje Ritsma zich voor een delicaat olympisch vraagstuk geplaatst. (*Vk* 12–11–01)

I replaced the absolute *with*-construction by a relative clause with a tensed verb and a subject, because the acceptability of the variants with the *with*-construction is influenced by the ease of interpretation of the implied agens of *gericht*. This factor seems to me to be different from the complexity factors discussed in this paper.

2. We analysed the positions of complex and simple constituents in five journalese texts: three soccer reports and two columns. The total of clauses analysed is 181. See Jansen and Wijnands (to appear).

3. Hawkins (1998) is interesting because he applies his processing approach to Danish, which is also a V2 language. Hawkins analyses the Danish equivalents of (1) as root sentences with the fronted constituent and the remainder of the sentence as daughters. Hawkins is able to demonstrate that the fronted constituents in Danish are in most cases shorter than the remainder of the sentence, which is in accordance with the predictions of EIC. Hawkins does not discuss the effects of constituents in first position in relation to the same constituents in other positions.

4. Hawkins (p.c.) is currently working on a revision of his EIC-theory in which complexity is considered an aggregation of the recognition costs of all the relevant syntactic and lexical domains of the sentence (see Hawkins 2001 for lexical domains). When applied to Dutch, this revised theory gave promising results.

5. This generalization does not hold for the subjects of presentative clauses, where the subject occupies positions at the end of the sentence.

References

Behaghel, O. (1909) Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25, 110–142.
Gibson, E. (1998) Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68, 1–76.
Gibson, E. (2000) The dependency locality theory: a distance based theory of linguistic complexity. In: Marantz, A., Miyashita, Y. & W. O'Neil (eds.) *Image, Language, Brain*. Cambridge (Mass): The MIT-press. p. 95–126.
Hawkins, J.A (1994) *A performance theory of order and constituency*. Cambridge: Cambridge University Press.
Hawkins, J. A. (1998) A processing approach to word order in Danish. *Acta Linguistica Hafniensa*, 30, 63–101.
Hawkins, J. A. (2001) Why are categories adjacent? *Journal of Linguistics* 37, 1–34.
Jansen, F. & Wijnands, R. (to appear) Doorkruisingen van het LinksRechtsprincipe. *Neerlandistiek*.
Maureau, J. H. (1983[2]) *Goed en Begrijpelijk schrijven*. Muiderberg: Coutinho.
Renkema, J. (1981) *De taal van Den Haag*. Den Haag: Staatsuitgeverij.
Renkema, J. (1995) *Schrijfwijzer*. Den Haag: SdU.