

# Centering and Scrambling: Towards a Pragmatic Motivation for Russian Word Order.

Sophia Malamud

smalamud@babel.ling.upenn.edu

My many thanks go to Ellen Prince, without whose prodding and generous advice this paper would not have been written. I am also grateful to Eleni Miltsakaki for help with LaTeX and generous discussions.

## **1 Introduction**

A definition of scrambling already poses a problem in a study that shuns commitment to a syntactic framework. I shall be using the term to denote, vaguely, a process by which two grammatical clauses differing only in the order of their constituents may be formed in a language.

A number of works investigating the phenomenon have appeared in the

past decade, mostly in the generative tradition (Mahajan 1990, King 1993, Bošković and Takahashi 1995, Miyagawa 1997, and many others). The main problem that has been noted for these analyses was the apparent lack of motivation for this type of movement. In fact, for most cases of scrambling, no purely syntactic or semantic motivation could be found (Kondrashova 1997, Bošković and Takahashi 1995, King 1993, *inter alia*). Thus, we turn to pragmatics (focus structure; other aspects of information packaging, discourse salience) for motivations of scrambling (e.g. King 1993, Kondrashova 1997, Bailyn 2000).

This paper considers a possible pragmatic motivation for ordering of noun phrases in Russian narrative. In particular, it argues that the noun phrases may scramble so as to allow for a smoother transition between clauses within a discourse segment. Centering Theory (Brennan, Friedman, and Pollard 1987, Walker, Iida and Cote 1994) allows us to formalise the notions of "smoothness" and "transition," based on the salience of discourse entities.

The hypothesis set forth in this paper is that Russian scrambling directly affects the smoothness of inter-sentential transitions, and that its motivation is to effect a smoother transition than would otherwise be produced.

To test this hypothesis, I performed a Centering analysis of narrative discourse segments in four large books, a short story, and an essay. The works were selected to represent a variety of narrative styles, and to contain

a sizeable number of narrative segments. Simple past-tense narrative was chosen as the discourse type that is least affected by considerations of poetic style - that is, the breaking of the pragmatic norms to achieve an artistic effect.

The results of this study show that discourse salience is indeed affected by word order in Russian and in turn, affects the smoothness of inter-sentential transitions. Therefore, producing a more coherent discourse arises as a motivation for scrambling.

This paper is organised as follows: in sections 2 and 3 I present the background information on Russian word order and the Centering theory, respectively. Theoretical discussion specific to this study follows in section 4. Section 5 consists of the description of data and analysis in this study, with discussion and conclusions drawn in section 6. Finally, the references and some examples of data analysis follow.

## **2 Background: Russian Word Order.**

Russian is a "free" word order language. This means that the word order in Russian does not encode "who did what to whom." Instead, the word order and suprasegmental phonology are used to encode different pragmatic and grammatical factors in an utterance.

In a Russian simple transitive "John killed Mary" theoretically all six permutations of the words yield grammatical sentences. However, a random

association of a permutation with a discourse context typically produces infelicity. Intonation contours also constrain the use of different word orders.

In her monumental 1986 study of Russian word order, Olga Yokoyama argues that the mechanisms of encoding the pragmatic and grammatical information in an utterance depend most heavily on the "speaker's subjective evaluation of the discourse situation" (Yokoyama 1986, p.331). That is, the word order and suprasegmental phonology that are used in Russian to encode this information depend on the speaker's assessment of the "knowledge" and "current concern" sets for those involved in the discourse.

Yokoyama's study was concerned with the spoken discourse, and her most definite conclusions dealt with Type 1 ("neutral") intonation contour for the utterances. Her work has partially formalised the Prague school's "theme-rheme condition" (Lenerz 1977, p.63), which suggests that "an NP may scramble over less rhematic NP" (non-literal translation in Rambow 1993)

The multiplicity of factors affecting Russian word order have been explored in a very principled way by Yokoyama. It seems, nevertheless, to be an exceptionally difficult task to formalise the process of "subjective evaluation of the discourse situation." The need for such formal description has been the motivation for this study. Centering Theory is a formal framework that allows to capture the above notions.

### 3 Background: the Centering theory.

#### 3.1 The Centering Transitions

The Centering Theory has been proposed in Grosz, Joshi, and Weinstein 1983 as a model that accounted for the use of different types of referring expressions. The ideas were subsequently developed and expanded both by the original authors, and by others (Brennan, Friedman, and Pollard 1987; Walker, Iida, and Cote 1994; Walker, Joshi, and Prince 1998, inter alia).

The Centering Algorithm allows us to compute the smoothness of transition between utterances based on salience ranking of entities in a discourse.

**Definition 1:** In each utterance, the set of discourse entities evoked in it is the set of *forward-looking centers* (Cf). Centers are semantic entities that are part of the discourse model (see Heim 1983), or items in the set of shared current concern (Yokoyama 1986).

**Definition 2:** There is a special member of this set called the *backward-looking center* (Cb). This is the entity that is most central in the utterance (Walker and Prince 1994), the file card you're writing on (Heim 1983), approximately corresponding to "the utterance theme" (Reinhart 1981, Horn 1986). The Cb links the current utterance with the previous discourse.

The set of forward-looking centers is ranked according to discourse salience, or "activatedness". The factors that determine ranking are the crux of the Centering Algorithm. If any centers are evoked in the next utterance, the highest-ranked of them is the Cb of that utterance. In fact,

if there is a pronoun in an utterance, then the Cb of this utterance is also denoted by a pronoun.

**Definition 3:** The highest-ranked center is the *preferred center* (Cp). It predicts what the next utterance is going to be about.

The interaction between Cb and Cp determines smoothness of transition from one utterance to the next as shown in the table 1 below. When the most central entity in an utterance ( $Cb(U_n)$ ) is the same as the most central entity in the previous one ( $Cb(U_{n-1})$ ), and the same item is also predicted to be central in the next utterance ( $Cp(U_n)$ ), the resulting discourse is very coherent, and the transition is *Continue*. On the other hand, when the Cb from a previous utterance is retained as such, but not predicted to be as salient in the next utterance, the transition type is *Retain*. The two Shifts result when the most central entity changes: the *Smooth-Shift* predicts that it should not change again in the next utterance, while the *Rough-Shift* does (Table 1).

**Table 1.** Transitions from  $U_{n-1}$  to  $U_n$ .

	$Cb(U_n) = Cb(U_{n-1})$	$Cb(U_n) \neq Cb(U_{n-1})$
$Cb(U_n) = Cp(U_n)$	Continue	Smooth-Shift
$Cb(U_n) \neq Cp(U_n)$	Retain	Rough-Shift

$U_n$  -  $n^{th}$  utterance,  $Cb(U_n)$  - the backward-looking center of the  $n^{th}$  utterance,  $Cp(U_n)$  - the preferred center of the  $n^{th}$  utterance.

The transitions, smoothest to roughest are, thus: Continue, Retain, Smooth-shift, and Rough-shift (Walker and Prince 1994). Centering analyses have shown that smoother transitions are preferred over rougher ones within a discourse segment (Di Eugenio 1990, Rambow 1993, *inter alia*).

### 3.2 The Ranking

The ranking of entities determines the Cp of the current utterance, and the Cb of the next one. The ranking principle arrived at by most Centering analyses (e.g. Di Eugenio 1998, Miltsakaki 1999), is based on the grammatical function of the entities, which are ranked as follows: EMPATHY → SUBJECT → OBJECT → OTHER. Here, "empathy" denotes either phrases grammatically marked as "empathic" (e.g., in Japanese), or otherwise clearly emphasising the experiencer (e.g., in the dative subject constructions (Yokoyama 1986)).

Studies of Italian (Di Eugenio 1998), Turkish (Hoffman 1998), and Greek (Miltsakaki 1999) have shown that this ranking indeed correctly predicts full noun phrase, pronoun, or zero-pronoun usage, and is independent of the utterance word order in these languages.

However, a study of German (Rambow 1993) showed that whereas topicalisation interacts with Centering in an ambivalent way, scrambling directly affects the ranking: "the Cf (ordered set of forward-looking centers) of an utterance is the list of constituents of the *Mittelfeld* in that order."

## 4 The Centering study of Russian scrambling

The hypothesis of this study concerns precisely the ranking of discourse entities. When the order of noun phrases differs from the basic one (subject, object, other), I hypothesised that the ranking goes by word-order: left to right. Thus, a reordering of noun phrases may be prompted by the transition preference of the speaker or writer.

### 4.1 The segment, the utterance, and other ranking assumptions

Centering is a model of the local discourse structure: it operates within "discourse segments." Hence, it is important to know how to determine the segmentation of a discourse. However, determination of segment boundaries is a separate question of much current investigation. I therefore assume no a priori segmentation in written discourse.

Within each segment, the Centering algorithm calculates the Cf list for every "utterance" - another notion in need of formal definition. Early Centering analyses seem to assume the utterance to be approximately the tensed clause (Kameyama 1998). In a later investigation (Miltasakaki 1999), this was revised, and "utterance" was defined as a full sentence, i.e., "the main clause and its accompanying subordinate and adjunct clauses" (Miltasakaki and Kukich 2000). I follow here the revised definition. Miltasakaki 1999 argues that the ordering of subordinate and main clauses does not

affect Centering. In this study, therefore, unless there were two or more coordinated subordinate clauses, everything but the main clause was ignored in the utterance.

There has been much variation as to the correct ranking of entities within a complex noun phrase (e.g., possessive). I am following, rather arbitrarily, a left-to-right ranking convention for these phrases.

Otherwise, I have performed data analysis using two different rankings: first, by grammatical function within the main clause, and, subsequently by word order within the main clause.

## 5 Data and Results

### 5.1 Control Data

The main source of data in this study was the online library of Russian literature ([www.lib.ru](http://www.lib.ru)). To provide a measure of the true proportions of different transitions in Russian texts, a full short story "Pyat' minut vzajmy" was chosen and analysed. A total of about 70 transitions was calculated. Since of the 78 sentences containing 24 transitive clauses with overt arguments only 4 were scrambled, the ranking was performed by grammatical function only. Discounting the rough-shifts in the opening and closing paragraphs of the story as "necessities of artistic considerations," the analysis has shown that Rough-Shifts constituted 10% of all the transitions, with the remaining comprising 34 Continues, 16 Retains, and 13 Smooth-Shifts.

## 5.2 The Rough-Shift measure

In order to check the validity of the left-to-right ranking hypothesis, I have chosen a number of literary narrative segments containing scrambled sentences. A computerised search was used to select the segments containing scrambled sentences from electronic books. A total of 44 analysable segments of two or more sentences were found, each containing at least one scrambled sentence.

The Centering analysis of this data was done manually twice (see the Appendix). The first analysis utilised ranking by grammatical function and produced 50 Continues, 46 Retains, 16 Smooth-Shifts, and 17 Rough-Shifts out of 129 total transitions. Then, using the left-to-right ranking hypothesis, the second analysis was performed, producing 49 Continues, 46 Retains, 21 Smooth-Shifts, and 13 Rough-Shifts.

In their 2000 Centering study, Miltsakaki and Kukich argue that "in general, Continues, Retains, and Smooth-Shifts do not yield incoherent discourses" (Miltsakaki and Kukich 2000). Therefore, only the presence of a Rough-Shift signals a significant incoherence. The number of Rough-Shifts and the presence of Rough-Shifts in a perceptually coherent discourse were therefore the first considered factors in this study.

Statistical tests were then run on these numbers, with the transition percentages from the short story analysis serving as controls, e.g., the norm. Although the tests indicated that the second analysis was much closer to the

normal data, the sample was not large enough to yield a degree of certainty above 75 data was performed. A closer examination of the transitions has indicated that of the 17 Rough-Shifts produced by the first analysis, 6 were found to be Smooth-Shifts in the second. One of these could have been a Continue changing to Retain in the second analysis, depending on the judgement of the main clause boundaries. The remaining 11 were Rough-Shifts in the second analysis as well.

At the same time, out of the 13 Rough-Shifts produced by the second analysis, 2 were Smooth-Shifts in the first. One of these could have been actually a Smooth-Shift in the second analysis if a "mop kicking and striving for the window" could be considered animate. Overall, therefore, it seems that the word-order dependent ranking provides a more accurate measure of coherence for the data. Since ranking the subject higher shows more discourse segments to be incoherent, it is patent that an unscrambled sentence may produce an incoherence where a scrambled one would not.

### **5.3 New hypothesis: incorporating the verb**

The two analyses have produced approximately the same number of Continue and Retain transitions. Moreover, both analyses have "improved" and "worsened" about the same number of these transitions. This suggests that the original hypothesis does not sufficiently account for the more coherent data. A motivation for reorderings that result in the above transitions has to be found.

In the original hypothesis of this investigation, no consideration has been given to the verb. However, it has been noted for many languages, including Russian, that the pre-verbal and post-verbal positions in an utterance have very different informational functions (Yokoyama 1986, Rambow 1993, Kiss 2000, *inter alia*). Therefore, the position of the verb was traced in the 34 scrambled transitive sentences with overt arguments for which the two analyses give different transitions. For 15 of them the word-order dependent ranking (second analysis) produced a smoother transition, whereas for the remaining 19, the other analysis did.

Crucially, 12 of the former sentences had OVS order, whereas 16 of the latter had the order OSV. It becomes obvious, thus, that simply scrambling the object to the sentence-initial position in Russian doesn't affect its discourse salience, but serves some other purpose. When, however, the subject is simultaneously demoted to the post-verbal position, the salience of both entities is affected.

Of the remaining 3 sentences "confirming" the left-to-right ranking hypothesis, two were the only VOS utterances in the data, and one more depended on the possessor-possessee ranking. At the same time, of the remaining 3 sentences contradicting the original hypothesis, all were OVS. However, two of them were a part of the 6-utterance parallel construction segment, and one more a part of a segment in which calculation of segment boundaries and, therefore, of the Cbs was very difficult.

Thus, the new hypothesis is formulated as follows: **the entities in**

**Cf are ranked left-to-right in scrambled sentences, except when scrambling is limited to bringing the object to sentence-initial position.** This revised hypothesis was used for the third and final analysis of the data. The analysis produced only 11 Rough-Shifts, 20 Smooth-Shifts, 35 Retains, and 63 Continues. These are the smoothest resulting transitions yet. The chi-squared test was used to measure the probability that observed differences in all the variables (the number of occurrences of each transition type) are the result of chance variation. In this test, the significance of the Rough-Shift measure is somewhat downplayed, since each variable is given the same significance in the calculation of the chi-squared value. Again, the percentages from the short story analysis were used as controls. As is evident from table 4, the new hypothesis results in a significantly more normal analysis of the scrambling data (60% probability that the difference from normal is chance).

**Table 2.** The chi-squared test.

Ranking hypothesis	Chi-squared value	Probability
By word order	12.69	Less than 1%
By gram. function	16	Even less! (about 0%)
The new hypothesis	2.07	60%

Thus, indeed the scrambling that "topicalises" the object while leaving

subject-verb order intact does not affect the salience ranking of entities. In this, it patterns like the German topicalisation to the Vorfeld (Rambow 1993) with respect to Centering. At the same time, all the other types of scrambling in Russian behave exactly like the German scrambling of constituents in the Mittelfeld.

## 6 Conclusions.

The above result suggests that Centering and word order are interdependent phenomena. Thus, if one follows the main claims of the Centering Theory that inter-utterance transitions are ordered by preference (Continue→Retain→Smooth-shift→Rough-shift), and that a Rough-Shift signals a breach in coherence, then a motivation for reordering noun phrases arises.

A study of spoken narrative must be done to test the applicability of the above results. The data collection for such a study is indeed under way now. On the other hand, a formal model of Russian word order and information structure incorporating the above insights is needed. An attempt at such a model, in the framework of Set-Combinatory Categorical Grammar proposed for Russian in Nygren 1999, is currently in progress.

## References

- Bailyn, J. F. (2000). Does Russian Scrambling Exist? Ms. SUNY at Stony Brook, *International Conference on Word Order and Scrambling*, Tucson, AZ.
- Boškovič, Ž. and Takahashi, D. (1995). Scrambling and Last Resort. Ms. University of Connecticut and Graduate Study Center, CUNY.
- Brennan, S., Friedman, M. and Pollard, C. (1987). A Centering Approach to Pronouns. *Proc. 25th Annual Meeting of the ACL*, Stanford, CA.
- Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.
- Grosz, B., Joshi, A. and Weinstein, S. (1995). Centering: a framework for modelling local coherence of discourse. <http://www.cis.upenn.edu/ircs>.
- Di Eugenio, B. (1998). Centering in Italian. In M. Walker, A. Joshi, and E. Prince (1998)
- Freedman, D., Pisani, R., and Purves., R. (1998). *Statistics*. 3rd ed. New York: W.W. Norton & Company.
- Heim, I. (1983). File change semantics and the familiarity theory of definiteness. In R. Bauerle, C. Schwarze, and A. von Stechow (eds.). *Meaning, use and the interpretation of language*. Berlin: Walter de Gruyter.
- Hoffman, B. (1998). Word Order, Information Structure, and Centering in Turkish. In M. Walker, A. Joshi, and E. Prince (1998).
- Kameyama, M. (1998). Intrasentential Centering: A case study. In M.

- Walker, A. Joshi, and E. Prince (1998).
- King, T. (1993). Configuring Topic and Focus in Russian. Ph.D. Dissertation, Stanford University.
- Kiss, K. (2000). Movement to the left periphery. Ms. Linguistic Institute of the Hungarian Academy, *International Conference on Word Order and Scrambling*, Tuscon, AZ.
- Kondrashova, N. (1997). Generativnaja grammatika i problema svobodnogo poryadka slov [Generative Grammar and the Problem of Free Word Order]. In A. Kibrik, I. Kobozeva, and Sekerina I. (eds.). *Fundamental'nye napravleniya v sovremennoj amerikanskoj lingvistike. [Fundamental Trends of Modern American Linguistics.]* Moscow: MSU Press.
- Lenerz, J. (1977). *Zur Abfolge Nominaler Satzglieder im Deutschen*. Tübingen. As cited in Rambow 1993.
- Mahajan, A. (1994). *Toward a Unified Theory of Scrambling*. Berlin: Walter de Gruyter & Co.
- Miltsakaki, E. (1999). Dissociating Discourse Salience from Information Structure: Evidence from a Centering study in Modern Greek and Japanese. *Computational Linguistics in the Netherlands, (CLIN '99)*.
- Miltsakaki, E. and Kukich, K. (2000). Automated Evaluation of Coherence in Student Essays. To appear in *Proceedings of ACL 2000*. <http://www.ling.upenn.edu/elenimi/grad.html>
- Moshkow, M. On-line Russian Library. <http://www.lib.ru>.
- Nygren, N. (1999). Coordination and Word Order in Russian. Ms. Uni-

versity of Edinburgh, Edinburgh, UK.

Prince, E. (1981). Toward a Taxonomy of Given/New Information. In P. Cole (ed.). *Radical Pragmatics*. New York: Academic Press.

Radford, A. (1997). *Syntactic Theory and The Structure of English - A Minimalist Approach*. New York: Cambridge University Press.

Rambow, O. (1993). Pragmatic Aspects of Scrambling and Topicalization in German: A Centering Approach. Unpublished Manuscript, University of Pennsylvania.

Reinhart, T. (1976). The Syntactic Domain of Anaphora. Ph.D. dissertation, MIT.

Walker, M. and Prince, E. (1996). A Bilateral Approach to Givenness: a Hearer-Status Algorithm And a Centering Algorithm. In T. Fretheim, and J. Gundel (eds.). *Reference and referent accessibility*. Philadelphia: John Benjamins.

Walker, M., Iida, M. and Cote, S. (1994). Japanese Discourse And the Process of Centering. *Computational Linguistics 21*.

Walker, M., Joshi, A. and Prince, E. eds. (1998). *Centering Theory in Discourse*. Oxford: Oxford University Press.

Webber, B. (1991). Structure And Ostension in the Interpretation of Discourse Deixis.

<http://www.ling.upenn.edu/~ellen/bonnie.ps> .

Yokoyama, O. T. (1986). *Discourse and Word Order*. Philadelphia: John Benjamins.

## Appendix: a worked example

Consider the following segment from Bulgakov, in which the second sentence is scrambled (The C<sub>b</sub> of the previous utterance is K.):

- (1) K. svistnul.  
K. let-out-a-whistle.  
'K. let out a whistle.'

C<sub>f</sub> = {K.}

C<sub>p</sub> = K.

C<sub>b</sub> = K.

Etogo svista Margarita ne uslyhala, no ona ego uvidela v  
Of-this whistle Margarita not heard, but she it saw at  
to vremya, kak ee vmeste s goryachim konem brosilo  
that time, as her together with hot horse it-threw  
sazhenej na desyat' v storonu.  
sazhens for ten to side.

'Margarita didn't hear this whistle, but she saw it at the same time  
when she, together with her hot-tempered horse, was thrown several  
meters to the side.'

Analysis 1: ranking by grammatical function.

C<sub>f</sub>={Margarita, whistle, horse}

C<sub>p</sub>=Margarita

C<sub>b</sub> = whistle

Transition = Rough-Shift

Analysis 2: ranking by word order.

$C_f = \{\text{whistle, Margarita, horse}\}$

$C_p = \text{whistle}$

$C_b = \text{whistle}$

Transition = Smooth-Shift