Incorporating Discourse Aspects in English – Polish MT

MALGORZATA E. STYŚ & STEFAN S. ZEMKE
University of Cambridge & Linköping University

Abstract

English orders constituents in utterances according to their grammatical function. Polish places them with regard to their informational salience and stylistic criteria. This rises two problems when translating: how to determine informational salience and which potential order to prefer. The former is addressed by providing an extended version of the centering algorithm. The latter, by extracting order preferences from statistical data.

1 Introduction

Machine translation systems tend to concentrate on conveying the meaning and structure of individual sentences. However, since translation has to be accurate not only lexically and grammatically but also needs to carry across the contextual meaning of each utterance, incorporating discourse aspects is necessary.

English and Polish exhibit certain idiosyncratic features which impose different ways of expressing the information status of constituents. Unlike English, in which constituent order is grammatically determined, Polish displays an ordering tendency according to constituents' degree of salience, so that the most informationally salient elements are placed towards the end of the clause. Such ordering requires solid knowledge about the constituents' degree of salience.

This paper is organised as follows. The next section includes a description of the centering algorithm for English and our extensions of the notion in view of English – Polish machine translation. We then go on to describe the idiosyncratic properties of Polish and their implications for center transfer. Finally, the rules for ordering Polish constituents are outlined.

2 Centering model for English analysis

Centering as introduced by (Grosz et al. 1986) is a discourse model proposing rules for tracking down given information units on local discourse

level. Center, expressed as a noun phrase, is a pragmatic construct and it is intentionally defined as the discourse entity that the utterance is about.

Each utterance U is assigned a forward-looking center list $\mathrm{Cf}(U_n)$ of all nominal expressions within the utterance ordered by their grammatical function which corresponds to the linear order of constituents in English. The backward-looking center $\mathrm{Cb}(U_n)$, the center proper, is the highest ranked element of U_n which is also (if possible) realised in $\mathrm{Cf}(U_{n-1})$. Pronominalisation and subjecthood are the main criteria underlying this ranking. Generally, (resolvable) pronouns are the preferred center candidates. For possible relations between subsequent utterances look at (Brennan et al. 1987).

2.1 *Extension to the centering algorithm*

Various refinements have been added to the centering model since its introduction (Brennan et al. 1987; Kameyama et al. 1986; Mitkov 1994; Walker et al. 1994). Description of our practically motivated extensions follows.

2.1.1 Definiteness. Definite articles often point to a center. However, the correlation between definiteness and an entity having been introduced in previous discourse is high but not total. (For example, proper names can be textually new yet definite.) We therefore include definiteness among factors contributing to center evaluation. Indefinite noun phrases are treated as new discourse entities.

2.1.2 Lexical reiteration. Lexically reiterated items include repeated or synonymous noun phrases possibly preceded by articles, possessives or demonstratives. We also propose to consider semantic equivalence based on the synonyms coded in the lexicon as valid instances of reiteration.

2.1.3 Referential distance. For pronouns and reiterated nouns, we propose the allowed maximal referential distance, measured in the number of clauses scanned back, to correlate with the word length of the constituent involved (Siewierska 1993a). This relates to the observation that short referring expressions have their resolvents closer than longer ones. Such precaution limiting the referential distance minimises the danger of over-interpretation of common generic expression such as *it*. We have not yet experimented with various functions relating the type of referent to its allowed referential distance, a simple linear dependence (with factor 1-2) seems to be reasonable.

	CONSTRUCT	MARKERS	CENTER VALUE
	Center-pointing constructions (Point. 1-4)		
1	Cleft	it+Be+N_c+that/who	center(N_c):=3
2	Fronted	N_f,Sentence-N_f	center(N_f):=3
3	Prompted	Prompt+N_p,Sentence	center(N_p):=3
4	There-insertion	there+Be+N_t	center(N_t):=2
	Pronominal centers (Pron. 1-2)		
1	Personal	I/you/it/he/she/we/they	center($Pron_{pers}$):=2
2	Demonstrative	this/that/these/those	center($Pron_{demo}$):=1
	Other (Non. 1-3)		
1	Indefinites	a/an/another/other	center(N_{indef}):=-1
2	Proper names	e.g., Mary/Chicago	center(Proper):=1
3	Default for any NP	*Cases not listed elsewhere*	center(NP):=0
	Composite Centers (Comp. 1-5)		
1	Reiterated nominal	$N_{reit} \overset{ref\_dist}{\longleftrightarrow} N_{reit}$	center(N_{reit})+1
2	Definite expressions	the/such/this/that etc. +N	center(N)+1
3	Possessives	its/his/her etc. +N	center(N)+1
4	Genitives	N's+N_p, N_p+of+N	center(N_p)+center(N)
5	Resolved pronoun	$NP_{match} \overset{ref\_dist}{\longleftrightarrow}$ Pron	center(Pron)+1

Table 1: *Center values for different types of NP*

2.1.4 Center-pointing constructions. Certain English constructions unambiguously point to the center thus making more detailed analysis unnecessary. The cleft construction uses a dummy subject *it* to introduce center, e.g., *It was John who came.* The center can also be fronted, e.g., *Apples, Adam likes*, or introduced by a prompt *as for, concerning, with regard to* etc., e.g., *As for Adam, he doesn't like apples.*

2.1.5 Composite center value. The rules for Composite Centers in Table 1 allow us to calculate center value increase over the default 0. Thus, for example, the center value for *the scientists' colleagues* will be arrived at by adding the contribution for *the* (+1) to the contributions for *scientists* and *colleagues* (each 0 or 1 depending on whether the item is reiterated) giving a value between 1 and 3 depending on the context. A constituent is assumed to be assigned the highest possible center value allowed by our rules.

2.1.6 Center gradation. Considering the priority scale of referential items, the mechanisms underlying centering in English could then be outlined as follows,

- Preference of pronouns over full nouns.
- Preference of definites over indefinites.
- Preference of reiterated items over non-reiterated ones.
- Preference of constituents involving more 'givenness' indicators.

These considered along with special center-pointing constructions lead to the numerical guidelines presented in Table 1. Some of them agree with the idea of the givenness hierarchy cf. (Gundel 1993). In Table 3, we illustrate the application of rules included in Table 1.

We choose the constituent with the highest center value as the discrete center of an utterance. If more than one constituent has been assigned the same value, we take the entity that is highest-ranked according to the ranking introduced in the original algorithm (Grosz et al. 1986; Grosz et al. 1995; Brennan et al. 1987).

	UTTERANCE	RULES	VALUES	CENTER
1	The scientists	Comp.2	1 = 1+0	scientists
	conducted many tests.	Non.3	0	
2	The tests were thorough.	Comp.1,2	2 = 1+1+0	tests
3	The results were looked	Comp.2	1 = 1+0	
	at by their colleagues.	Comp.3,5	2 = 1+1+0	colleagues
4	They were acknowledged.	Pron.1, Comp.5	3 = 2+1	they
5	The scientists' colleagues	Comp.1,2,4	3 = 1+1+1+0	colleagues
	accepted the tests.	Comp.1,2	2 = 1+1+0	

Table 2: *Center values for example clauses*

3 Local discourse mechanisms in translation

In discourse analysis, we relate particular utterances to their linguistic and non-linguistic environment. Below, we shall describe the relationship

between the grammatical sentence pattern (Subject Verb Object) and the communicative pattern (Theme Transition Rheme).

Functional sentence perspective (FSP) is an approach used by the Prague School of linguists to analyse utterances of Slavic languages in terms of their information content (Firbas 1992). In a coherent discourse, the given or known information, *theme*, usually appears first thus forming a co-referential link with the preceding text. The new unit of information, *rheme*, provides some specification of the theme. It is the essential piece of information of the utterance.

There are clear linear effects of FSP[1]. Utterance non-final positions usually have given information interpretation whereas the final represents the new. This could be motivated by word order arranged in such a way that first come words referring to details already familiar from the preceding utterances/external context and only then words describing new detail. Similarly, in perception first comes identification and only then augmentation by details individually connected with the given idea (Szwedek 1976).

Constituent order in Polish generally follows the communicative order from given to new. Since the grammatical function is determined by inflection, there is great scope for the order to express contextual distinctions and the order often seems free due to virtual absence of structural obstacles. However, there are also other, mostly stylistic, factors influencing the final order which can co-specify or even override the 'given precedes new' tendency. This presents a delicate task of balancing a number of clues selecting the most justified choice. The degree of emphasis is also a factor and it is worth noting that the more frequently an order occurs the less emphatic it is (Siewierska 1993a).

4 Ordering of Polish constituents

Our choice of ordering criteria has been directly inspired by the findings of the Prague School discussed above, our own linguistic experience (both of us bilingual, native speakers of Polish) some statistical data provided by (Siewierska 1987; 1993a,b) and by the feasibility of implementation. The intended approach to ordering could be characterised as follows,

[1] The information structure also changes depending on the accentuation pattern, but we shall leave the intonation aspects aside in this presentation.

Permissive: Generate more (imperfect) versions rather than none at all. If need be, restrict by further filters.

Composite: Generate all plausible orders before some of them will be discriminated. (This approach is side-tracked when a special construction is encountered.)

Discrete: No gradings/probability measures are assigned to competing orders as to discriminate between them. This could be an extension.

4.1 *Ordering criteria*

Below we present some rules which are obeyed by Polish clauses under usual conditions:

- End weight principle: Last primary constituent is the anti-center;
- Given information fronting: Constituents belonging to the given information sequence are fronted;
- Short precedes long principle: Shorter constituents go first;
- Relative order principle: Certain partial orders are only compatible with specific patterns of constituents.

Additionally, there is a strong tendency to omit subject pronouns. Such omission, however, exhibits different degrees of optionality. What follows is a list of constructs used in subsequent tables to generate plausible orders of (translated) Polish constituents.

Center information: has the highest rank in the ordering procedure and is used in three aspects:

- center(*Constituent*) returns the center value of the *Constituent*'s NP, or 0 if undefined,
- center_shift(*Utterance*) holds if *Utterance* relates to the preceding one in the way allowed by the shift transition cf. (Grosz et al. 1986a)
- discrete_center(*Constituent*) holds if *Constituent* is the chosen center of the current utterance

Length of constituents: length(*Constituent*) returns the number of words of the resulting Polish *Constituent*[2]. Although not as important as center information, this rough measure can discriminate certain orders on the basis of 'short precedes long principle'[3].

[2] This measure, to a great extent, depends on the translation of constituents. It could be approximated by the length of the original English, instead of Polish, units. We use that in the example.

[3] However, for the otherwise rare order OSV, the opposite applies.

Positioning of certain constituents: (or indeed their lack) can in turn induce other constituents to occupy certain positions. Some orders are only possible in certain configurations, e.g., with frontal *Adjunct* (X-), whereas others require just its presence (-X-), or absence (X=[ ]).

Syntactic phenomena:

- grammatical function of a constituent, e.g., being a subject (S) or object (O).
- pron(S) & pron(O) if both subject and object are pronominal or $\mathrm{Sub}(U_n) = \mathrm{Sub}(U_{n-1})$ – if subject stays the same.
- certain expressions, e.g., a focus binding expression such as 'only', can trigger specific translation patterns.

Features of next utterance: e.g., $\mathrm{center}(S,U_{n+1}) > 0$, can be used together with the features of the current utterance in order to obtain more specific conditions.

In the following tables S denotes (Polish) subject, V – verb, O – object, X – adjunct, Prim – S or O, "-" – (sequence of) any, [ ] – omitted constituent. The difference for "$\gg$" to hold must be at least 2.

4.2 *Building on orders of constituents*

The Preference Table ?? presents some of the main PREFERENCEs for generating orders of Polish constituents depending on specific CONDITIONS. Each line of the table can be treated as an independent if-then rule co-specifying (certain aspects of) an order. Different rules can be applied independently thus possibly better determining a given order[4]. The JUSTIFICATION column provides some explanation of the validity of each rule.

It might be the case that as a result of applying the Preference table, we obtain too many orders. The Discrimination Table 4 provides some rationale for excluding those matching ORDERs for which one of their DISCRIMINATION conditions fails. If the building stage left us with no possible orders at all, we could allow any order and pick only those which successfully pass all their discrimination tests. It is purposeful that all orders apart from the canonical SVO have some discrimination conditions attached to them. The rarer the order tends to be the more strict the condition. Therefore, SVO is expected to prevail. Both the Preference table and the Discrimination table are mostly based on statistical data described in (Siewierska 1987; 1993a,b).

There remains a number of cases which escape simple characterisation in terms of 'preferred and not-discriminated'. The Preprocessing Table 5

[4] Orders derived by co-operation of several rules could be preferred in some way.

Pref.	CONDITIONS	PREFERENCE	JUSTIFICATION
	Orderings implied by center information		
i	center(Any) < 0	-Any	Final position of new
ii	center(C1) $\gg$ center(A2)	-Any1-Any2-	Given-new principle
iii	center(X) > 1	X-	Adjunct topic fronted
iiib	discrete_center(Prim)	(X-)(V-)Prim-	Primary center fronted
	Statistical positioning preferences		
iv	-V-S-O- & -X-	XV-S-O-	Statistical (66%)
v	-O-S- & X-	XV-O-S-	Statistical
vi	-V-O-S- & -X-	XV-O-S-	Statistical (53%)
vii	-S-V-O- & -X-	XS-V-O-	Statistical (32%)
viii	-S-V-O- & -X-	S-V-OX	Statistical (30%)
ix	-O-V-S- & -X-	O-V-SX	Statistical (29%)
x	-O-V-S- & -X-	O-VXS	Statistical (26%)
xi	Pron(S) (& center_shift(U_n))	-VS-	Stylistic
xii	*General*	-V-O-	Statistical (89%+)
xiii	*preferences*	-S-O-	Statistical (81%)

Table 3: *Center values for example clauses*

offers some solutions under such circumstances. It is to be checked for its conditions before any of the previous tables are involved. If a condition holds, its result (e.g., 0-anaphora) should be noted and only then the other tables applied to co-specify features of the translation as described above. The Preprocessing table can yield erroneous results when applied repeatedly for the same clause. Therefore, unlike the other tables, it should be used only once per utterance.

In Table 6 we continue the example from Table 3. The orderings built on by a cooperation of the Preprocessing/Preference and not refused by the Discrimination table appear in the last column.

5 Conclusion

One of the aims of this research was to exploit the notion of center in Polish and put it forward in context of machine translation. The fact that centers are conceptualised and coded differently in Polish and English has clear repercussions in the process of translation. Through exploring the pragmatic, semantic and syntactic conditions underlying the organisation

Discr.	ORDER	DISCRIMINATION	JUSTIFICATION
i	-V-S-O-	length(S) $\leq$ length(O)	Statistical (99%)
ii	-V-S-O-	-V-S-O	Statistical (87%)
iii	-V-S-O-	Pron(S)	Stylistic
iv	-V-O-S-	length(O) $\leq$ length(S)	Statistical (96%)
v	-V-O-S-	-X- present	Statistical (89%)
vi	-S-O-V-	SOV	Statistical (50%+)
vii	-S-O-V-	center(S,U_{n+1}) > 0	Statistical
viii	-O-S-V-	OSVX	Statistical (79%)
ix	-O-S-V-	length(O) $\geq$ length(S)	Statistical (100%)
x	-O-V-S	length(O) $\geq$ length(S)	Statistical (64%)

Table 4: *Discrimination table*

Pre.	CONDITIONS	RESULT	JUSTIFICATION
	0-anaphora		
i	S='we'	S=[ ]	Rhythmic
ii	pron(O) & pron(S)	S=[ ]	Stylistic
iii	Sub(U_n) = Sub(U_{n-1}) (& pron(S))	S=[ ]	Stylistic
iv	center_continuing(U_n)	S=[ ]	Stylistic
	Special constructions		
v	-'only' SV- & pron(S)	-'tylko' SV-	Focus binding expr.
vi	X=[ ] & pron(O)	SOV	Special: S,O,V only

Table 5: *Preprocessing table*

of utterances in both languages, we have been able to devise a set of rules for communicatively motivated ordering of Polish constituents.

Among the main factors determining this positioning are pronominalisation, lexical reiteration, definiteness, grammatical function and special centered constructions in the source language. Their degree of topicality is coded by the derived center values. Those along with additional factors, such as the length of the originating Polish constituents and the presence of adjuncts, are used to determine justifiable constituent order in the resulting Polish clauses.

	PREFERENCE CRITERIA	PARTIAL ORDERS	DISCRIMINATION (FAILING)	RESULTING ORDER(S)
1	Pref.xii	SVO		SVO
	Pref.xiii	VSO	(Discr.iii)	
2	*No rules apply, order unchanged*			SVX
3	Pref.iiib	OVS	Discr.x	OVS
	(Pref.xii)	VOS	(Discr.v)	
		OSV	(Discr.viii)	
4	Pre.iii	S=[ ]		V[S]X
	Pref.xi	-VS-		
5	Pref.iiib	SVO		SVO
	(Pref.xii)	VSO	(Discr.i)	

Table 6: *Example continued: Deriving constituent orders*

In further research, we wish to extend the scope of translated constructions to di-transitives and passives. We shall also give due attention to relative clauses. Centering in English can be further refined by allowing verbal and adjectival centers as well as by determining anti-center constructs.

We have thus tackled the question of information distribution in terms of communicative functions and examined its influence on the syntactic structure of the source and target utterances. How and why intersentential relations are to be transmitted across the two languages remains an intricate question, but we believe to have partially contributed to the solution of this problem.

REFERENCES

Brennan, Susan E., Marilyn W. Friedman & Carl J. Pollard. 1987. "A Centering Approach to Pronouns". *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL'87)*, 155-162. Stanford, Calif.

Firbas, Jan. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge: Cambridge University Press.

Grosz, Barbara J. 1986. "The Representation and Use of Focus in a System for Understanding Dialogs". *Readings in Natural Language Processing* ed. by Grosz, Barbara, K. Jones & B. Webber, 353-362. Los Altos, Calif.: Morgan Kaufmann Publishers.

____, Aravind K. Joshi & Scott Weinstein. 1995. "Centering: A Framework for Modelling the Local Coherence of Discourse". *Computational Linguistics* 21:2.203-225.

Gundel, Jeanette K. 1993. "Centering and the Givenness Hierarchy: A Proposed Synthesis". *Workshop on Centering Theory in Naturally Occurring Discourses.* Philadelphia: University of Pennsylvania.

Kameyama, Megumi. 1986. "A Property Sharing Constraint in Centering". *Proceedings of the 24th Annual Conference of the Association for Computational Linguistics (ACL'86)*, 200-206. Columbia, N.Y.

Mitkov, Ruslan. 1994. "A New Approach for Tracking Center". *Proceedings of the International Conference "New Methods in Language Processing"*, 150-154. Manchester: UMIST.

Siewierska, Anna. 1987. "Postverbal Subject Pronouns in Polish in the Light of Topic Continuity and the Topic/Focus Distinction". *Getting One's Words into Line* ed. by J. Nuyts and G. de Schutter, 147-161. Dordrecht: Foris.

____. 1993a. "Subject and Object Order in Written Polish: Some Statistical Data". *Folia Linguistica* 27:1/2.147-169.

____. 1993b. "Syntactic Weight vs. Information Structure and Word Order Variation in Polish". *Journal of Linguistics* 29:233-265.

Szwedek, Aleksander J. 1976. *Word Order, Sentence Stress and Reference in English and Polish.* Edmonton: Linguistic Research, Inc.

Walker, Marilyn A., Masayo Ida & S. Cote. 1994. "Japanese Discourse and the Process of Centering". *Computational Linguistics* 20:2.193-227.