World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:8, 2007

Continuous Text Translation Using Text Modeling in the Thetos System

Nina Suszczanska, Przemyslaw Szmal, and Slawomir Kulikow

***Abstract*—**In the paper a method of modeling text for Polish is discussed. The method is aimed at transforming continuous input text into a text consisting of sentences in so called canonical form, whose characteristic is, among others, a complete structure as well as no anaphora or ellipses. The transformation is lossless as to the content of text being transformed. The modeling method has been worked out for the needs of the Thetos system, which translates Polish written texts into the Polish sign language. We believe that the method can be also used in various applications that deal with the natural language, e.g. in a text summary generator for Polish.

***Keywords*—**anaphora, machine translation, NLP, sign language, text syntax.

I. Introduction

ALGORITHMS discussed in this paper have been worked out for the needs of the Thetos system, which translates texts in Polish into the Polish sign language. The translation process as well as problems connected with it were discussed e.g. in [1], [2]. Thetos involves a linguistic analysis subsystem, whose task is to transform the input text into a sequence of words that are correct sentences according to the sign language grammar. Those words are signed, one by one, by a separate animation subsystem. As a result of that, we get on the screen an animated sequence of gestures shown by a virtual character specially designed for that purpose [3].

A natural requirement for translation is to preserve the content conveyed by the input text. In order to achieve that, in our system the translation (transfer) is done after a semantic analysis, preceded by a morphologic and syntactic one. During semantic analysis, a semantic representation is built for each input sentence. It is then transformed into a corresponding structure of one or more output language sentences. At last for each of them a surface structure is generated according to the sign language grammar. Let's note that in order to make translation easier, during the syntactic analysis compound input sentences are automatically splitted into component ones.

Previously, in our works, inter-sentence links existing in the continuous text were not considered during transformation. Such a translation method has proved not to be sufficiently good [4]. In order to improve the translation results for continuous texts, we have proposed a method which is based on modeling the input text before it is translated.

Manuscript received November 20, 2004. This work was supported in part by the Polish Committee for Scientific Research in 2003-2005 under Grant 4 T11C 024 24.

N. Suszczanska, P. Szmal, and S. Kulikow are with the Institute of Informatics, Silesian University of Technology, 44-100 Gliwice, Poland (P. Szmal's phone: +48-32-237-2883; fax +48-32-237-2733; e-mails: <Nina.Suszczanska, Przemyslaw.Szmal, Slawomir.Kulikow>@polsl.pl).

II. Continuous Text Organization. Inter-Sentence Links

One of functions of texts is to convey some content. To express a complex content we use longer, multi-sentence continuous text. Between sentences material links appear; they are connected with repetitions of information. In case where for interpretation of some expression in the text we need information provided by its antecedent, we have to do with anaphora.

Links are mirrored in the construction of the text; the construction conforms to definite stylistic and pragmatic rules. One of them is so called economy rule, what means that the structural scheme of a sentence is not fully embodied in it. Some information repetitions have no corresponding verbalization in the surface structure of the sentence. One kind of such non-verbalized abbreviations in sequential syntactic structures is substitution of some fragment of the structure with a (lexical) zero, or – in other words – ellipsis.

We proposed a notion of canonical form for sentences. A sentence is in such form if it is a single one, has a predicative center, and in its structure defined by the valence scheme of the predicate, all obligatory places are filled up. It means that a sentence in a canonical form does not contain such structural abbreviations as ellipses and anaphoras.

Sentences that compose a continuous text can be transformed into the canonical form during the text modeling process. The presence of anaphoras seriously complicates this task. A problem appears how to correctly substitute specific words with terms which have an equivalent sense. Another problem is how to transform elliptic sentences in which we have to recognize lacking elements and to reconstruct appropriate terms. In both cases analysis of inter-sentence links is necessary [4].

During processing of ellipses an assumption holds that: a) substitution of one word with another which has the same sense does not change the sense of entire utterance, and b) substitution of one word with another which has the same denotate (i.e. corresponding notion in the real word) does not change the denotate of the primary construction. Additional complications arise when in repetitions (verbalized or not) one term is changed with another which has a close denotate with another meaning, but this changes neither the denotate nor the sense of the utterance. For example: *dziewczynka (girlie),*

Czerwony Kapturek (Little Red Riding-Hood), grzeczna dziewczynka (good girlie); the terms differ in gender and denote different types of objects. Examination of such phenomena exceeds the scope of the semantic analysis and should involve an analyzer which refers to text pragmatics.

Due to some essential features of the sign language, sentences in the canonical form are worth considering in the context of translation into this language. First of all, it is symptomatic of literality. Its syntax is closely connected with its semantics – syntactic structure of sentences is subordinate to the logical one. Order of words is strictly defined: in the first place the agent always stands, followed by the predicate, then – objects, at last – possibly adverbials. Moreover, the language has no but simple sentences (there is no compound sentences in it).

The problem of analyzing text syntax is very complex, has many aspects, and in this moment it has no unambiguous solution, even a theoretical one. In our approach the analysis covers only selected set of syntactical phenomena, and we are foreseeing extension of this set. To our experiments we took first of all fairy tales, because we expected links between sentences to be clearer, comprehensible for young children. The text examples in this paper come from the fairy tale about the Little Red Riding Hood. On these examples the reader can see that the texts of tales abound with inter-sentence links, too.

III. Organization of Translation in the Thetos System

A. Translation scheme

The method used in the Thetos system to translate text into the sign language can be represented with a scheme shown in Fig. 1. The full scheme (inclusive of both light- and dark-shadowed blocks) is relative to the present version; the scheme limited to light-shadowed blocks answers the primary version of the system. By the way, a technical detail: input and output data of all processors are written using the XML language.

In extension of the description given in the Introduction, the reader should note that linguistic processing is divided into two parts. In the first one, a model of the input text is built – individual sentences are transformed to canonical form. In the second part, they are translated into a text conformed to the sign language grammar rules. In the next subsections we will shortly discuss selected stages of processing.

B. Syntactic analysis

The syntactic analysis is done in the Polsyn parser. It consists in grouping words according to the rules of the syntactic groups grammar for Polish (SGGP) and in detecting relations that occur between syntactic groups (SG) on all levels. SGs are characterized by their type, a unique name (identifier), and grammatical features. The syntactic representation of a sentence is a labeled graph whose nodes are SG's and edges – relations between them. The root of the graph is a verb SG (VG); graph topology corresponds to the conditions imposed by the SGGP. SG's can enter into relations with other SG's; a relation that occurs between the VG and some other SG defines some syntactic role in the sentence.

The way the parser works will be demonstrated on an example. The text to be processed is shown in Fig. 2. In Fig. 3 abbreviated results of syntactic analysis are shown; to get the full results one has to run the linguistic analyzer within the linguistic server LAS: http://thetos.zo.iinf.polsl.gliwice.pl/las/.

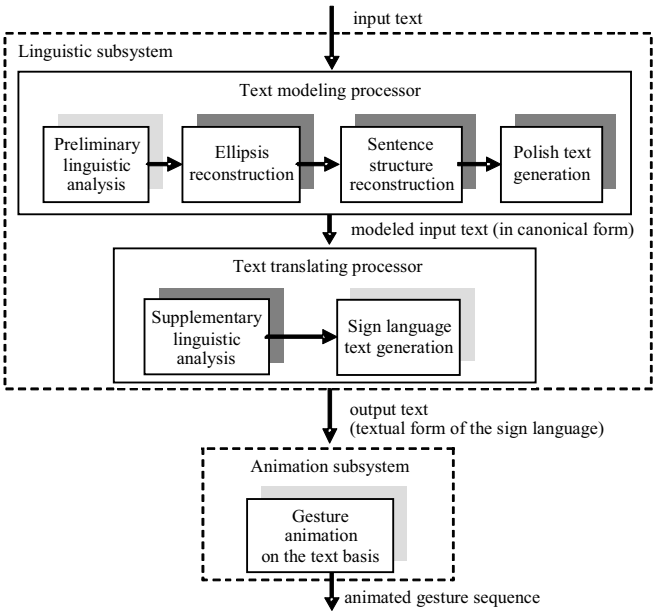

FIG. 1. TRANSLATION SCHEME IN THE THETOS SYSTEM
(DARK-SHADOWED BLOCKS DID NOT APPEAR IN THE PRIMARY VERSION)

Sentence	Polish text	English equivalent
	<input_text>	
S1	Dziewczynka chodziła w czerwonej pelerynce z kapturkiem	The girlie wore (litt.: *walked in*) a red pelerine with a hood
S2	i dlatego wszyscy nazywali ją Czerwonym Kapturkiem.	and therefore everyone called her the Little Red Riding Hood.
S3	Jej mamusia także lubiła używać tego imienia,	Also her mother liked to use that name
S4	bo pasowało do dziewczynki.	since it fitted to the girlie.
	</input_text>	

FIG. 2. FRAGMENT OF THE LITTLE RED RIDING HOOD FAIRY TALE
(ORIGINALLY LF'S DO NOT APPEAR IN THE TEXT)

Both sentences appearing in the sample text are compound. During the syntactic analysis each of them was split on two, in effect we obtained four simple sentences (S1 – S4). For each of them individually a parse was done. The parser found, among others, that S2 and S3 contained anaphoras, and in S4 the subject was omitted. In the figure, NG, PG, and AG denote noun, prepositional, and another SG respectively.

C. Preliminary semantic analysis

The result of the semantic analysis is a predicate-argument (P-A) structure. In general case it consists of a predicate and nine arguments. It takes the form: *Predicate*(a_1, …, a_9).

Arguments a_1-a_9 have the following semantic labels: *Agent, Object, Recipient, Instrument, Location, Time, Goal, Cause, Another*. In case, when the predicate has no specific

arguments or some arguments do not appear in the sentence, respective entries in the structure have zero (0) value.

Sentence	Recognized attributes
S1	+ <group type="**S**" id="**S1**" level="**0**" childCount="**5**" children="**NG11,PG4,PG5,AG7,VG8,**" mainChild="**VG8**" mainChildPos="**5**" kind="**1,-**" function="**major**" features="**-**" relations="**SynR=#predicate(VG8,S1), SynR=#subject(NG11,VG8), SynR1=#adverbial_gr(PG4,VG8),SynR2=#adverbial_gr(PG5,VG8),**" isLeaf="**0**" lineNumber="**1**">
S2	+ <group type="**S**" id="**S2**" level="**0**" childCount="**6**" children="**AG8,NG14,NG15,NG16,AG9,VG9,**" mainChild="**VG9**" mainChildPos="**6**" kind="**4,5**" function="**sentence**" features="**-**" relations="**SynR=#attr(AG8,VG3), SynR=#predicate(VG9,S2), SynR1=#obj4(NG15,VG9),SynR1=#obj5(NG16,VG9),**" isLeaf="**0**" lineNumber="**24**">
S3	<group type="**S**" id="**S3**" level="**0**" childCount="**8**" children="**NG17,AG10,NG18,{AG11,AG12},NG19,AG13,VG10,**" mainChild="**VG10**" mainChildPos="**8**" kind="**1,3,2**" function="**major**" features="**-**" relations="**SynR=#attr(AG11,VG6),SynR=#predicate(VG10,S3), SynR=#subject(NG18,VG10), SynR1=#obj3(NG17,VG10),SynR1=#obj2(NG19,VG10),**" isLeaf="**0**" lineNumber="**47**">
S4	<group type="**S**" id="**S4**" level="**0**" childCount="**4**" children="**AG14,PG6,AG15,VG11,**" mainChild="**VG11**" mainChildPos="**4**" kind="**-**" function="**sentence**" features="**-**" relations="**SynR=#attr(AG14,VG7),SynR=#predicate(VG11,S4), SynR1=#adverbial_gr(PG6,VG11),**" isLeaf="**0**" lineNumber="**78**">

FIG. 3. RESULTS OF SYNTACTIC ANALYSIS (IN XML)

The P-A structure is built during semantic interpretation of SG's that participate in the parse and their syntactic roles. The interpretation may bring an ambiguous result. In order to limit complexity of further processing, we allowed the algorithm to examine at most four different parse variants. Moreover we allow only one parse variant to be translated; in consequence we have to unambiguously select the "best" variant first.

At the preliminary analysis stage a transitory (syntactic-semantic) structure is built, in which semantic interpretation is applied to NG's that pretend to the semantic roles of *Agent* and *Object*; other SG's are written into proper places labeled with syntactic roles. The look of the transitory representation of the sample text considered in Section III.B is shown in Fig. 4. In this specific case, there is no different parse variants in the structure; if they were, the section of the specific sentence in the table would be extended with 1, 2, or 3 rows.

Sentence	Predicate	Subject	Object 2	Object 3	Object 5	Adverb.1	Adverb.2
S1	VG8	NG11				PG4	PG5
S2	VG9	NG14	NG15		NG16		
S3	VG10	NG18	NG19	NG17			
S4	VG11					PG6	

FIG. 4. RESULTS OF SYNTACTIC & SEMANTIC ANALYSIS

D. Reconstruction of ellipses

The objective of the next processing stage is to analyze the structure of the sentence and to supplement it so as it could be transformed into a canonical form, the content of the primary sentence being preserved. Repeated analysis covers also those SGs to which the roles *Agent* and *Object* have been assigned.

The transformation algorithm may be described as follows:

1) Analyze the structure of VG. If VG is null (what occurs when there is no predicate in the sentence), activate J. Romaniuk's algorithm to supplement the null substitute [4].
2) On the base of VG's structure, generate a generalized *Predicate* and put it into proper entry of the P-A structure.
3) In the dictionary Semsyn [5] select for *Predicate* the scheme which matches best the parse of the sentence.
4) Using the scheme found in Semsyn, check if NG has been correctly put to the *Agent* position. If there is no agent (or subject, on the syntactical level), activate the algorithm for its reconstruction with appropriate personal pronoun.
5) Put the selected NG at *Agent* entry in the P-A structure.
6) Repeat actions 4) and 5) for the role *Object*. If there is no object in the sentence while one appears in the predicate's scheme, activate object reconstruction algorithm using the appropriate pronoun.
7) On the basis of the scheme from Semsyn dictionary, find and put remaining SGs from the sentence parse into respective P-A structure entries. To do that, perform semantic interpretation of respective SGs.

While searching a „hidden" SG we assume that processing is being done in the framework of the same continuous text. Processing methods have been only implemented for some specific ellipsis types. They are some cases of so called "conjunction reduction", where in one of two sub-sentences making up a compound one, shared predicate or – in some case – subject or object is omitted. In case of zero-substitution of lacking predicate we assume that it is always in the preceding sentence. In our implementation we used some J. Romaniuk's algorithms for substitution of text that contains ellipses. We decided to temporarily use heuristics in place of some too complex algorithms.

As an example can serve the two-part sentence (S3-S4 from Fig. 2):

*Jej mamusia także lubiła używać tego **imienia**, bo pasowało [...] do dziewczynki. (Her mother also liked to use that **name** since [it] fitted to the girlie.)*

In the place marked with "[...]" the pronoun "ono" ("it") is omitted; it stands there for the word "imię" ("name").

In effect of recovery performed by using some heuristic, the sentence S4 took the form:

*Bo **ono** pasowało do dziewczynki. (Since **it** fitted to the girlie.)*

This form is fully acceptable; the change in the word order has no influence on the content of the sentence.

E. Supplementing the structure of the sentence

At the stage of supplementing the P-A structure, final text recovery takes place. At this stage, we consciously break the language economy rules and replace anaphoras with their antecedents.

The problem can be stated as follows: find an anaphora in the text, retrieve the NG which is its counterpart an put it in the anaphora's place. The text is viewed as a known space within which we can move forward and backward in search

after inter-sentence links. There are several types of anaphoric links; we implemented one of them: reference of a pronoun to an NG. Now we can only handle retrospective references, where the antecedent (that is the NG being searched) appears in the text *before* the pronoun. It also holds for the pronouns used to supplement the text at the previous processing stages.

Algorithm for generating pronoun→NG links is based on structural-valence schemata from the Semsyn dictionary. At first the deep structure of the sentence is built. In the structure all obligatory places in the verb's valence scheme should be filled-out. If we ascertain that some places stay not filled, we try to find candidates for these places in preceding sentences.

Before we started elaboration of algorithms, we made a statistical exploration of what the distance is of the reference which is sought. We used texts accessible in the Internet (articles in IT periodicals, belles-lettres, fables for children). It proved that the distance of research of the reference can be constrained to three sentences, including the current one.

F. Surface structure generation

Generation of the surface structure of the text being modeled takes place after anaphora analysis is done. In this case, like it is during building and filling up the deep structure, the valence scheme from the Semsyn dictionary is helpful.

A sample of results given by the anaphora analyzer is shown in Fig. 5. In the text model, with further processing in sight, the analyzer – instead of the intended inflective form of respective word – inserts its basic form plus a morphological description of the very form.

Sentence	Polish text	English equivalent
	<xmlText>	
	...	
S3	**{dziewczynka:1:221231261} mamusia także lubiła używać tego imienia** .</simpleSentence>	{girlie: ...} mother liked to use that name .
S4	<simpleSentence id="**S4**"> **bo {imię:1:311} pasowało do dziewczynki** .</simpleSentence>	since {name: ...} fitted to the girlie.
	</xmlText>	

FIG. 5. RESULT OF MODELED TEXT LINEARIZATION

IV. PROBLEMS OF LINGUISTIC PROCESSING

To start with linguistic processing problems: processing is now divided into two parts. At first, a model of the input text is built – individual sentences are transformed to the canonical form. Obtained sentences are then translated into text conforming to the sign language grammar rules. The algorithm can be improved by removing the stages of linearization and repeated analysis of the modeled text. After that, the results of the linguistic analysis would be directly edited. We are working at an instrumentarium which would enable us to safely do this job.

There still remain a lot of serious problems connected to continuous text analysis. We are planning to solve some of them in the future: proper reconstruction of missing predicate, recovery of anaphoras expressed in the text by means of synonyms and "cognated" terms, which appear in the same conceptual hierarchy. Let's expand the idea. Reach resources of the natural language allows us to express the same content with different wordings, and in consequence – different SGs. Semantic representation of the sentence is dependent of its syntactic representation. Therefore sentences that are conveying the same content while having different syntactic structure, may differ in semantic representation. A task of the semantic analysis could be detecting the equivalence of sentences; after transformation, the same set of sentences in canonical form should be obtained (or sets equivalent in the sense of content being conveyed). It will allow to get more precise translations than they have been hitherto.

In the latter task it would be profitable to use a semantic representation in the form of a set of binary semantic relations between arguments. On such sets, reasoning rules could be introduced that should allow us to reconstruct missing relations that hold between notions in easier and more naturally than it is in case of the P-A representation.

At the end we'd like to signal a problem which is important for translations into the sign language. The language contains non-verbalized elements used to define space and time [6]. Now we are only gathering ideas how to cope with that.

V. CONCLUSION

In the paper a method of generating a computer model of texts in Polish was discussed. The effect of introducing such a model into the process of translation into the sign language was – first of all – improvement of results of generation of sign language utterances in a textual form. It is much easier to manipulate modeled text units than units of the primary text. The method for text modeling was developed for the purposes of the Thetos system that translates Polish text into the Polish sign language. However, the text model can be also used in other applications that have to do with natural language processing, e.g. in text summary generation. On the basis of the model it is easier to build a semantic representation of the text, in form of binary semantic relations between arguments of a predicate-argument structure.

REFERENCES

[1] P. Szmal, N. Suszczańska, "Selected problems of translation from the Polish written language to the sign language", *Archiwum Informatyki Teoretycznej i Stosowanej*, vol. 13, no.1, pp. 37-51, 2001

[2] N. Suszczańska, P. Szmal, J. Francik, "Translating Polish Texts into Sign Language in the TGT System", in *Proc. 20th IASTED Int. Conf. Applied Informatics AI'2002*. Innsbruck, Austria 2002. pp. 282-287.

[3] J. Francik, P. Fabian, "Animating Sign Language in the Real Time", in *Proc. 20th IASTED Int. Multi-Conf. Applied Informatics AI 2002*, Innsbruck, Austria, 2002, pp. 276-281.

[4] S. Kulików, J. Romaniuk, N. Suszczańska. "A syntactical analysis of anaphora in the Polsyn parser". in *Proc. Int. IIS:IIPWM'04 Conf.*, Zakopane, Poland 2004, pp. 444–448.

[5] Grund, D. "Computer implementation of syntactical-generative Polish verb dictionary (Komputerowa implementacja słownika syntaktyczno-generatywnego czasowników polskich)". *Studia Informatica*, vol. 21, no. 3(41), pp. 243-256, 2000.

[6] M. Swidzinski, T. Galkowski, *Study on lingual competence and communication of deaf (Studia nad kompetencją językową i komunikacją niesłyszących)*. Warsaw: Warsaw University, ISBN 83-904863-2-6, 2003.