COMPLEX PROCEDURES FOR MT QUALITY

Michael Zarechnak
Georgetown University

The purpose of this paper is to propose methods for the improvement of existing working MT by blazing a path to the Multi-Language Intermediate Language (MLIL). The goal can be reached by the use of effective tools. It is not expected that the road to this new plateau will be an easy one to travel. Yet one wants to believe that the travel is feasible and rational. A typical working MT system as we know it is built around a pair of languages. These systems are reaching the point of diminishing return since the primary "progress" relies on the adding of "clustered entries" into the dictionary without any insightful semantic coding to facilitate the degree of formalization with subsequent algorithms based on it. What is the solution? People's activities should give us a clue. In translators' continuing struggle with translation problems across languages and cultures plus the societal sets of values and communication tools, we observe the very consistent attempt to grasp the situation reflected in the source message. By selecting some focal point in that situation, translators strive to express it in the appropriate equivalence, composed of the form, meaning, function and other components necessary to make it reasonably acceptable to the target language audience. The user should be the final judge. The symbiosis between the output from MT and the consumer is the solution to this difficulty. We see the whole process for MT as consisting of four elements: situation, source language, IL, target language. The specific stepping stones will lead us to:
1. The set of rules relating the real extralinguistic situation to binary codes in the dictionary and the formal rules based on these codes. 2. To make the feasibility study more effective, we shall limit the frame situations to a set of specific subfields so the control functions might become more secure.

MOTTO: NUR DAS BEISPIEL FUEHRT ZUM LICHT (VIELES REDEN TUT ES NICHT).

COMPLEX PROCEDURES FOR MT QUALITY

The purpose of this paper is to discuss five components which could be considered as useful in improving the quality of MT processes, whether on a pair or multi-language basis. The following components are listed;

1. Output from the GU Russian to English MT illustrating the need for its further improvement.
2. Exhaustive representation of the relationships between the form of a linguistic unit (F) and its meaning (M).
3. Hjelmslevian dependencies for syntagmatic functional types.
4. Categories of Aristotle adjusted for the semantic coding of dictionary entries.
5. An illustration of binary coded dictionary entries from weather report data.

The author believes that a successful MT system should eventually evolve into some variation of a fuzzy system in which the formal segments will be augmented by fuzzy logic complements. (See Bibliography for reference to Zadeh).

The MT output examined relates to coding errors, morphological description, syntagmatic classification, semantic roles and theoretical insights.

The fuzzy approach to translation is based on the fact that MT output definitely shows that human intuition, on one hand, is not formally captured, as is the case with the majority of examples given, but on the other hand due to the systematic treatment of structures, can sometimes replicate spontaneous human output, as is the case with the following sentence from a Russian scientific text;

(1) Naibolee soverwennym byl by takol diagnoz, kotoryl otrajal by vse perecislennye storony progressa.

(1.1) the human translation: The most perfect would be such a diagnosis which would reflect all sides of the process enumerated above.

(1.2) the MT system generation: Most complete would be such a diagnosis which would reflect all of the enumerated sides of the process.

The reader can see the facsimile output in Appendix #1.

The fact that the two outputs are very similar leads us to the assertion that the translation process produces dynamic equivalences characterized by inequalities rather than equalities in formal terms. This fact makes it necessary to formalize the possible arrangements between form (F) and meaning (M) based on the symbols of equality (=), approximation (~~) and inequality (#). The approximation relationship in translation could be compared with the degree of comparison characteristic of qualitative adjectives and also with Hjelmslev's dependencies. The net result of this kind of comparison is the assumption that ideally any morpheme is subject to vertical (inherent) and horizontal (syntagmatically relational) properties. It is also important to note that the ratio between the depth of coding for either of these two axes is mutually complementary; given more coding to the vertical (paradigmatic) axis of the morpheme, one can give less to the syntagmatic axis.

The next principal question, should the paradigmatic information be considered as nearly context-free and the syntagmatic information as nearly context-sensitive?

It is desirable that the representation of information be as economical as is feasible. One way of approximating economy is to code in bits 0 vs 1. Given a tree of generic information vs diagnostic and specific information, one can build a vector of binary codes such that each position in that vector will have a value of its own; presence vs absences of a particular value (8. p.68-69) Let us illustrate. We want to code a word such as ANIMAL, BIRD, HUMAN BEING, FISH, PLANT. What is common to all of them in terms of componential analysis is ABILITY TO MOVE. In terms of diagnostics, MOVE can be differentiated as by one's own will and ability or not. Thus BIRD, HUMAN, and FISH can be separated from PLANT. Specific aspects could be captured as follows; HUMAN +speech, BIRD +fly, FISH +swim. When we look at the words MOVE, SWIM, FLY and SPEAK we notice immediately that these verbs not only express certain actions, but also classify the objects which are capable of performing these acts.

ACTION is one of the Aristotelian categories (7. p.1-29). Let us list all of them. 1. Substance; such as horse, man 2. Quantity; four-footed 3. Quality; white 4. Relative; larger, half 5. Where; here, in the Lyceum, in the marketplace 6. When; yesterday 7. Posture; is sitting 8. Possession; has shoes on 9. Doing; cutting, burning, moving, speaking 10. Being affected by; being cut, being burned, flown, spoken.

Since in categorial grammar the truth function is relevant, one should be aware that in the Aristotelian division of concepts into 'combined' vs 'just by itself' only the combined unit permits affirmation or denial. Aristotle held that every uncombined expression signifies (denotes) at least one of the ten categories. It is clear that due to the combinatory nature and the distributional probabilities, these categories could be used as underpinning factors in Valency Theory and Dependency or Case Grammar versions.

How can the Aristotelian categories be translated into codeable features? In Lyons' discussion of determiners, quantifiers and classifiers we are reminded that in order to use the referring expressions correctly we might be forced to answer these questions; what is there?, which one?, how much?, what kind of?. Answers to these questions could provide the values for vertical and/or horizontal coding as in the expression 'sunny weather'. See Appendix #2. When one comes to the linear arrangement of text units, one should remember that essentially three operations are possible; word order within the same length, expansion of the length by insertion of some items, reduction of the length by discarding some units from the text. If the word order is the same, one can replace some units as in 'The seat of the table is hard' vs 'The seat of the chair is hard' where only one of them has the truth function.

Let me begin with a few examples from the GU MT Russian to English translation system where certain mistakes are to be corrected:

A. General Problems (high frequency, seemingly chronic)

1. Unnecessary usage of BY when occurring in a prepositional phrase; MEJDU NIT6H *I* QILINDROM /between the thread and BY the cylinder.
2. In the instrumental case with the conjunctional expansion; VPERVYE PRIMENENNY1 CERMAKOM I GERMANOM/first time applied by C. and BY G.
3. BY is used instead of WITH; DLITEL6NOST6H should read WITH A LENGTH OF, not BY LENGTH.
4. CEREZ produces in the English text either ACROSS or THROUGH but not properly so; CEREZ 5EL6/THROUGH A CRACK vs ACROSS A CRACK.
5. Consistent incorrect translation of UDAR as BLOW, in Physics, IMPACT is more appropriate, e.g. CEREZ PR4MYE UDARY/*THROUGH STRAIGHT BLOWS should read THROUGH DIRECT IMPACTS, PRI LOBOVOM UDARE/UPON FRONTAL BLOW/IMPACT.
6. Problem with variation in translating PUT6; CISTO REZONANSNY1 PUT6/PURELY RESONANT *WAY/PATH
7. Some outstanding examples of the apparent word order problem are: CERTO1 OBOZNACENO USREDNENIE PO VREMENI/* BY CHARACTERISTIC WAS INDICATED THE NEUTRALIZATION FOR TIME, could possibly read INDICATING WITH A LINE THE NEUTRALIZATION ACCORDING TO TIME.
8. UMEN6WAETS4 BOL6WE/DECREASED *GREATER could read IS FURTHER DECREASED.
9. PRI STREMLENII VOZMOJNO TOCNEE MODELIROVAT6 USLOVI4 V ... /UPON STRIVING POSSIBLY *ACCURATE TO MODEL CONDITIONS IN ... , could read IN STRIVING TO SIMULATE MORE ACCURATELY CONDITIONS IN ...
10. V NASTO45EM EGO VIDE/IN PRESENT ITS FORM, should read IN ITS PRESENT FORM.
11. V SVETE RAZVITYX K NASTO45EMU VREMENI PREDSTAVLENI1/IN LIGHT OF IDEAS DEVELOPED TO THE PRESENT *TENSE OF PRESENTATIONS, could read IN LIGHT OF IDEAS DEVELOPED TO THE PRESENT TIME.
12. XOROWO PODVERTJDAETS4/WELL IS CONFIRMED, should read IS WELL CONFIRMED.

B. Sporadic Problems

1. S DRUGO1 STORONY-/WITH THE FRIEND OF SIDE should read ON THE OTHER SIDE.
2. K SOJALENIH-/TO REGRET, should read UNFORTUNATELY.
3. MOJNO VOSPOL6ZOVAT6S4/IT IS POSSIBLE TO BE EMPLOYED, should read IT IS POSSIBLE TO USE.
4. STALKIVAH5IXS4 CASTIQ/OF *PUSHED PARTICLES, should read OF COLLIDED PARTICLES.

It is clear that the MT output needs improvements on a variety of levels. How can we increase the quality? One solution is to keep comparing the massive output from the massive input with the view of diagnosing the problems and proposing general solutions not only for these problems but for problems of this type. Comparison will supply us with a large body of data which will permit sound generalizations between a pair of languages and a set of languages. Obviously, inductive observations alone will not suffice. We have to look for some insights based on these observations and our intuition in order to propose some formulaic representations of the data and its manipulation. The goal of a universal coding for the dictionary and a universal grammar for parsing and synthesis of the source language vs target language (SL vs TL) should be developed.

When we compare the output with the input, we find again and again that inequality rather than equality is obtained. Thus, assuming that form (F) vs meaning (M) should be compared, we may construct a matrix of all possible combinations to register which possibilities exist or are probable and which do not exist between a pair of languages. Let us compare Polish vs Russian. Since we know that equality will not be the overriding pattern, we shall introduce the relationships of approximation and inequality. The following formulaic expressions are used: = stands for equality, ~~ for approximation, # for inequality.

	Polish	Russian		

1.1	voda	voda	F = F
			M = M
				both forms and meanings are the same, E: WATER

1.2	chas	chas	F = F M ~~M	the form here is given in phonetic repres. cf. Polish CZAS, meaning approximates; P:TIME R: HOUR
1.3	kachka	kachka	F = F M # M	P: DUCK vs R: PITCHING (on the sea)
2.1	nieswiezy	nesvezij	F ~~F M = M	P: NOT FRESH, R: NOT FRESH

2.2	caly	tselyj	F ~~F M ~~M	P: ALL vs R: THE WHOLE an example: TO JEST CALY MOJ ZAPAS CUKRU/THIS IS THE WHOLE/ALL SUPPLY OF MY SUGAR
2.3	bilina	bylina	F ~~F M # M	P: A PLANT R: AN OLD RUSSIAN EPIC SONG

3.1	prawo	zakon	F # F
			M = M
				P: THE LAW, R: THE LAW

3.2	kilka	neskol'ko	F # F M ~~M	P: SEVERAL, R: SOME
3.3	falda	skladka	F # F M # M	P: WRINKLE, R: LOADING

When we move from the dictionary representations to the syntagmatic strings we soon discover that the Hjelmslevian (9. p.334) concepts of units within single syntagmas and crossing the length of single syntagmas becomes a real problem. This is true even when we stay on the morphosyntactic level, without going deeper into semantic relations. Discontinued complex morphemes occur and we have to look upon them as if they were contiguous and continuous units. Let us illustrate with examples from the process taking place in the comparison of the degrees applicable to adjectives and adverbs. This process can take the form of synthetic and analytic adjustment. Thus, in Russian and English one can have both the -ER type and MORE + the stem type of comparative forms as in LONG + ER, MORE UNPLEASANT. Given the three degrees of comparison; positive, comparative and superlative , one has to look for the set of forms which will approximate the meanings of those forms among a pair of languages and a set of languages.

Hjelmslev's dependencies are;

a) one-sided, determination (government, rection),
two-sided, reciprocal, interdependence (agreement, concordance) and compatibility dependence, constellation (adjoining), represented as a-->b, a<-->b, a---b respectively.

Comparison as a morphosyntactic process has a nexus, (i.e. nexial direction) is an exocentric structure and is heteronexial such that three instantiations are possible and all of them occur. Possibilities are;

(1) the left and right boundary occur at the polar ends of the string, e.g. 'She is as sweet as she could be.' 'The closer we try to get to the horizon, the further away it seems to be.'

(2) continuous display of the synthetic form as in 'He is taller than me.' 'He is more intelligent than me.'

(3) nondiscontinuous as in 'realizing his progress on a wider and wider scale.'

When we contrast comparison in a set of languages, we shall find out that determination and constellation will prevail in the majority of them. This presumed result should be adequate justification for coding any linguistic entry as to its participation in the comparison process.

We shall also refer to the computational approach of Zadeh (1. p.1-58) to fuzzy quantifiers which he bases on fuzzy logic and distributional semantics. In this semantics, the possibility vs probability is determined by a series of tests relating to the distribution of correlated linguistic units. One would see that the distribution of HOT vs TEMPER and WARM vs PERSONALITY is very revealing.

The design strategy would derive its force from the organic mixture of certain Aristotelian and Platonic concepts found later in other philosophical works such as Leibnitz, Locke, Lossky, Kholodovich, Zolotova, Dolinina and others.

It is my belief that a dictionary driven MT could serve as the basis for discovering what stepping stones should be followed along the path toward an Intermediary Language for multiple MT.

We think of the dictionary architecture essentially as a hierarchical cascade based on Aristotelian categories taken as a set of predicates for classification of dictionary entries.

```
NAIBOLEE                          ,4                  433P
SOVERWENNYM   S     5  P   3      ,3 5                433P
BYL                               *26  PA  P  MS    P4    CD 5 DXC
BY                                ,7                P4
TAKOI         S1    4  P          ,3 1                             A
DIAGNOZ       S1    4  P          *1 1                   CD2       A
*                                 /0
                                            APPENDIX #1
------ CUT ------
KOTORYI       S1    4  P          ,3 1
OTRAJAL                           *26  PA  A  MS    P4      CD Y5   G
BY                                ,7                P4              G
VSE           S1    4  P1   4     ,354                   H11        GA
PERECI'ENNYE  S        P1   4     ,230 PA  P                CD 5    GA
STORONY       S        P    4     $1 0                      cD 2    GAG
PROQESSA      S 2      P          /1 1                      cD 2     G
------ CUT ------
```

MOST COMPLETE WOULD BE SUCH A DIAGNOSIS , WHICH WOULD REFLECT ALL OF THE ENUMERATED SIDES OF THE PROCESS .

P* NAIBOLEE SOVERWENNYM BYL BY TAKOI DIAGNOZ , KOTORYI OTRAJAL BY VSE PERECISLENNYE STORONY PROQESSA *

APPENDIX #2.

EXPLANATORY NOTES ON INTRODUCTORY INPUT CODING FOR SEMANTIC BASE FROM WHICH THE SYNTACTIC FUNCTIONS COULD BE PROJECTED

Corpus: Weather Prediction

The input sentence: SUNNY AND WARMER WEATHER WILL OCCUR ALONG THE PACIFIC COAST.

Purpose:

To illustrate the procedure of how an entry from such a sentence is coded, what these codes stand for, and how at least a pair of words has to be coded in order to apply a SEMANTIC COMPATIBILITY RULE (SCR) to project the function on the syntactic level from the semantic base.

The context is considered as the field in which the syntagmatic properties and/or relations are manifested, not determined. Briefly: Each entry consists of the paradigmatic codes shown in the first row, Roman I with 16 binary positions, and the syntagmatic codes displayed in the second row, Roman II also with 16 binary positions filled with zeros or ones (0, 1). An example:

ENTRY	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
SUNNY ROMAN I	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	WORD 1 (W1),I
SUNNY ROMAN II	1	0	0	0	1	0	0	0	1	1	1	1	0	0	1	0	WORD 1 (W1),II
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
WEATHER1 ROMAN I	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	WORD 2 (W2),I
WEATHER ROMAN II	1	0	0	0	1	1	0	0	1	1	1	1	1	0	1	0	WORD 2 (W2),II

The interpretation of these codes are as follows:

positions I/1-4: 0010 5-8 9-12 13-16
abstract qualifier light natural temporal
attribute for
Word 1 (Wl),I

II/1-4: 1000 5-8
one- 'OBJECT'
place to be taken from the pos. 1-4 of
the W2, pos. 1-4,1 in our case:
WEATHER I/1-4:1000 ('object')
for WORD 1 (Wl), I

9-12 LIGHT:/natural
(look for it in the macro-, mezzo- and

micro-coding of word *2*, pos. 9-12), in our case the word WEATHER has a code in pos. 9-12:0001 for CLOUDINESS/LIGHT. This code is of temporal attribute kind. for WORD 1 (W1),I

13-16 We expect that the syntagmatic compatibility rule will be such that the head (0, WEATHER) will semantically include (A, SUNNY) as its qualifier resulting in an endocentric semantic relation coded as projected agreement dependence, i.e. 0010. for WORD 1(W1),I

VERBALIZATION

In word 1 we are saying that the entry SUNNY is an abstract class, a qualifier, with the semantic core 'LIGHT' as a terminal element for resonance test versus the second word, i.e. WEATHER, within which we would expect to find 'LIGHT' if these two words should be considered as 'semantically resonating'.

ENDO-relation via O vs A/Rhm
EXO-relation via O vs Rht/E.

IN OUR CASE SUNNY WEATHER WILL OCCUR we will have only ENDO-A, and EXO-E. But we should eventually consider the whole sentence:
SUNNY AND WARMER WEATHER WILL OCCUR ALONG THE PACIFIC COAST
and then all four types of structures will be recognized:

```
SUNNY AND WARMER WEATHER as endo-O/A/Rhm/A
  A   Rhm  A      O
WEATHER WILL OCCUR as Exo-0/E/E where EE is an identity
  O     E    E                                expansion
WEATHER WILL OCCUR ALONG THE PACIFIC COAST.
                     Rht       A      O as the exo-O/Rht
                                          including the
                                          endo-O/A.
```

(where A = attribute, O = object, Rhm = relator-homogeneous, Rht = relator-heterogeneous, E = event)

This ends the interpretation of Roman I for word 2: WEATHER. Syntactic projection function for the WEATHER: a possible SUBJECT/OBJECT of the SENTENCE depending on the SCR and special rules for disambiguation of SUBJECT vs. OBJECT syntactic functions.

The interpretation for W2 Roman II of WEATHER:

pos. 1-4: 1000 stands for one-position in exo-type, E.

5-8: 0100 finds this code in word z in position 1-4 coded as 0100/Roman I if not in position 1-4, then in pos. 5-8 of word y Roman II, Thus, in our case OCCUR will carry 0100 in pos. 1-4 Roman I and WILL will carry in pos. 5-8, Roman II, 0100.

9-12: A test for the semantic resonance between the

pair of the exo-relation words: WEATHER WILL
WEATHER (WILL) OCCUR
We generally expect that the texts constructed by the weather reports use only E that are compatible with the Os occurring in the same sentence.

13-16: 1010 means that the WEATHER governs WILL OCCUR and agrees with SUNNY: government between O vs E has a code 1000 agreement between O and A has a code 0010, which results in the code 1010. for word 2, Roman II.

Since the E is one place, there is no need for any syntactic object, accordingly the A Rhm A O E E Rht A O i.e. SUNNY (AND WARMER) WEATHER WILL OCCUR ALONG THE PACIFIC COAST, will be recognized as the O having the function of the subject in this sentence, and the A Rhm A as its attributes through endo structure, and Rht AO as the adverbial modifier of place.

SCR #1: IF THERE IS JUST ONE * PLACE E AND THERE IS JUST ONE O NOT INCLUDED IN THE Rht (A)O, THEN THIS O IS THE SUBJECT OF THAT SENTENCE. (The above rule disambiguates an O projected for Subject/Object syntactic function).
Since the ALONG PACIFIC COAST is included in Rht AO, COAST
O
carrying the code O is excluded from the subject function. Since the E both WILL and OCCUR are one-place Es, the only O still not taken is WEATHER, and hence it is defined as the SUBJECT in this particular sentence.

The above rule should serve only as an illustration to grasp the basic heuristic procedures in formulating in a very general way the rules for the syntactic functions to be projected form the semantic base.

REFERENCES

1. Aristotle's Categories and Propositions (De Interpretatione), Translated with commentaries and glossary by Hippocrates G. Apostle, The Peripatic Press Inc., 1980, p. 1-29.

2. Francis P. Dinneen, An Introduction to General Linguistics, Holt, Rinehart and Winston Inc. N.Y., 1967, p. 334.

3. A.A. Kholodovich, "On Typology of Word Order" in: Philological Sciences, No. 3, 1966 p. 3-13 (in Russian)

4. Eugene Nida and Charles Taber, The Theory and Practice of Translation, E.J. Brill, Leiden, Netherlands, 1969

5. L.A. Zadeh, "PRUF - a Meaning Representation Language for Natural Languages" in: Fuzzy Reasoning and its Applications. Ed. E.H. Mamdavi and B.R. Gaines, Academic Press, 1981, p. 1-58.

6. Michael Zarechnak and Edward Coyne, "Semantic Analysis of Natural Language Statements" in: Linguistics 182, Mouton Publishers, 1976, p. 73-81.

7. Michael Zarechnak, "The History of Machine Translation", in: Trends in Linguistics: Studies and Monographs, 11:, Mouton Publishers, 1979, p. 3-87.

8. Michael Zarechnak, "Outline of Semantic Research for Future MT Development" in: International Forum on Information and Documentation, vol. 5,2, Moscow 1980, p. 12-14.

9. Michael Zarechnak, "The Intermediary Language for Multilanguage Translation" in: Computers and Translation I, Paradigm Press Inc., 1986, p. 83-91.