[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland; ed. Kirsten Falkedal (Geneva: ISSCO).]

Recipes for Escaping from the Partial Ordering of Candidate Translations:* Some Consequences for the Evaluation of MT Systems

Jan Dings
EUROTRA-B / Katholieke Universiteit Leuven

Abstract

This paper deals with the question of whether abstract adequacy levels for Machine Translation systems can be formulated. The concept of partially ordered set is introduced to describe the phenomenon that candidate translations often seem to differ along the lines of mutually incomparable dimensions. The central idea is that MT systems can be ranked in terms of abstract adequacy levels, in function of how well they cope with the problem of partially ordered candidate translations.

1 The notion of adequacy levels

When I started my quest for abstract adequacy levels, it was brought to my attention that I should look for inspiration in the classical Generative Grammar framework, more precisely Chomsky '57 and '65. This would anchor my theory, and the resulting criteria, into a school of linguistic theory that cares about its meta-theoretical and epistemological status.

In these two works, Chomsky defines three "adequacy levels" for formal grammars; according to this hierarchy, a grammar G can be:

- OBSERVATIONALLY ADEQUATE: G produces the right set of sentences, i.e. exactly those sentences of language L that are grammatical according to the linguistic intuitions of a native speaker of L.
- DESCRIPTIVELY ADEQUATE: idem, but for each sentence, G produces the right structural analysis or derivation.

* This work has been supported by the Commission of the European Communities through the Eurotra Project.

- EXPLICATIVELY ADEQUATE: idem, but G is the best possible grammar, which means that it is as much integrated in UG (Universal Grammar) as possible. In Generative Grammar, a formulation of a given phenomenon in terms of linguistically universal concepts has the status of an explanation; to say that all languages share some feature, is giving the ultimate explanation for it: in such a case that feature is considered to be innate, which explains how a child can master it so quickly and with so little information.

The idea would then be to define a (human or mechanical) translational system as some kind of grammar, i.e. an intensional definition of some endless set, just like a grammar for a given language.

Take Gr(S) and Gr(T), the grammars for Source and Target languages respectively; Gr(S) and Gr(T) are generative definitions of the sentences of those two languages; i.e. a type of intensional definitions for the elements of an endless set.

Analogously an S-T translation system can be thought of as a grammar, as it is the intensional definition of the endless set of couples (s, S'), wherein s is a source language sentence, and S' the set of target language sentences that are "good translations" of s.

A quick and dirty analogy with the three adequacy levels above would then lead to the following:

- Observationally adequate: S-T produces the right set of translations for all sentences of the source language.
- Descriptively adequate: idem, but for each couple (s, S'), S-T produces a suitable structural description of how the translations have been obtained.
- Explicatively adequate: idem, but S-T is the best possible translation system between the languages involved, as it is formulated in terms of a general translational theory.

2 Adequacy levels for translations

The question is however: can I define a theory that provides for these three adequacy levels in the context of translation systems?
I believe that there are some problems that make this far from evident, such that it might be better to look elsewhere.

- Generative Grammar is based on the competence of the ideal speaker-hearer; its adequacy criteria focus the Language Acquisition Device of a child. Are there then ideal bilingual speaker-hearers (or reader-writers), capable of giving good-bad judgments for candidate (s, S') couples?
Judgments of this kind are crucially different from grammaticality judgments: it is a basic GG tenet that UG, the theory a grammar should be integrated with, is innate. And: whatever Universal Translation Theory might be developed some day, it seems we can be sure in advance that it will not be an **innate human capacity.**[1]

- It is clearly the case that good-bad judgments for translations are much more problematic than grammaticality judgments: especially between unrelated languages, the set of perfect (s, S') couples is very restricted, and in any case much smaller than the set of "acceptable" translations that are accepted **faute de mieux.** These "acceptable" (instead of **good,** or **the)** translations can be the subject of debates, and the boundaries of their set are much more vague than those of the set of grammatical sentences of a language[2]. As all theoretical argumentations and counter-examples that (try to) prove the impossibility of translation focus on the theoretical level of perfect translation, a theory of translation that takes into account the possibility of what is done every day, i.e. the possibility of actual translation practice, should speak in terms of **graded quality judgments,** instead of binary yes/no judgments (perfect translations). Otherwise we would speak about the translations of a minority of sentences and a minority of language-pairs.

It is ironic to note that current MT **systems,** on the one hand, work in a binary way: the theoretical assumptions they are based on force them to either find no translation at all, or a set (possibly a singleton) of sentences all considered to be perfect translations. **Evaluation** practice, on the other hand, seems to have a more flexible approach, using e.g. rating scales for fidelity, intelligibility, etc.

[1] In some sense, there is a general innate capacity involved in translating. Just as children can learn a specific language, on the basis of their knowledge of UG, people can learn to translate only because of ideas in sentences. The big difference is that virtually no one needs special training to learn to speak, while special training is needed even to make translators out of bilinguals.

[2] From a more strict theoretical point of view, the latter boundaries (i.e. of the set of sentences of la language) are crisp; it is the set of **acceptable** sentences that has vague boundaries, as linguistic **competence** if fuzzified by linguistic **performance,** and acceptability is performance-related (Cf. the number of embeddings, sentence length etc.)

From these ideas one can conclude that a realistic (machine) translation theory requires a device for quality **measurement,** instead of a binary criterion that distinguishes perfect translations from non-translations. Quality measurement however, turns out to be a very complex task, since it often implies the necessity of comparing different candidate translations along the lines of mutually incomparable dimensions. More concretely, if two translations have different degrees of imperfection along the lines of two or more mutually incomparable dimensions, we do not (yet) have the theoretical apparatus to explain or justify why one candidate translation is actually preferred by a **human translator,** given certain circumstances.

In more formal terms: the set of candidate translations is often a Partially Ordered Set, or **poset,** with several **maximal elements,** but without a **maximum.**

3 Algebraic intermezzo on posets

A set X is partially ordered by the relation R if

1. R is a subset of X x X (i.e. the cartesian product of X with itself).
   In words: R is a relation restricted to set X.
2. R is reflexive (i.e. for all x element of X, xRx holds).
3. R is antisymmetric (i.e. for all elements x and y of X, if xRy and yRx, then x=y).
4. R is transitive (i.e. for all elements x, y, z of X, if xRy, and yRz, then xRz).

Note that a poset need not be **linearly ordered** or **connected;** which means that there can be elements x, y of X, for which neither xRy nor yRx holds. A poset [X, R] is connected if for all x, y of X xRy or yRx holds.

In a poset [X, R] an element x is called **maximal** if for all z of X**,** if xRz, then x=z.

An element x is called the **maximum** of poset [X, R] if for all z of X, zRx holds.

In a poset [X, R] a maximum, if there is one, is maximal, and unique for X.

Example: if S = { {1}, {2}, {3}, {1,2}, {2,3} } then [S, $\subset$ [S] is a poset. (where $\subset$ [stands for the inclusion relation, restricted to the set S).

In this poset, {1,2} and {2,3} are maximal: they are contained in no other elements of S except themselves. There is no maximum.

There would be a maximum, however, if S contained also {1,2,3}.

In what follows, I will call MAX the set of maximal elements of a poset.

4 Dimensions for ranking translations

Now suppose we have a sentence or text fragment S in Source Language s. Ss is a complex constellation of form/content relations (where content is to be understood as total communicative value, a global term encompassing notions like **signifié, denotation, connotation, effect** etc.). Potentially every single form phenomenon (a word, its length and sound, its relative frequency, the syntactic rules it has been subjected to, etc.) can have its content counterpart; something which is exploited especially by literature. Think e.g. of poetry, where even the sounds of vowels and consonants all contribute to some "poetic" effect or impression in the mind of the reader. This is why only specialists (artists) of this particular kind of form/content relations, i.e. people who know (for both source and target language) what psychological effect words, their constructions, and their sounds may have upon their readers, can try to make acceptable translations of this kind of texts. In the extreme case of a **haiku** written in **kanji** by a calligrapher, translation would be impossible unless heavy loss of information is accepted. In ordinary language however, there is a large amount of redundancy, many form phenomena lacking a content counterpart. This implies that the same global content might be expressed in other ways (e.g. using other words, or other word order) and that there can be some amount of freedom for the translator.

In the Target Language then, it will not always be possible to find **the** perfect form counterpart for a given effect or content. Instead, several imperfect candidates can show up, all of them achieving different results for different content aspects. The translation of some word or phrase A can force the translator to make his choice between:

> B-l, having the same denotation, but coming from the wrong linguistic register (e.g. dutch **gangsterpraktijken → shady practices, unscrupulous practices.** Denotation is correct, but the style level is higher; dutch **futurum → future tense;** idem, even the sound is similar, but the style level is definitely lower).
>
> B-2, having the right denotation and style level, but having a

slightly different connotation, (e.g. dutch **echte gangsterpraktijken → the real Chicago touch;** denotation and style level are correct, but the connotation is radically different).

B-3, which has right denotation etcetera, but which is only one of the readings of a polyseme, such that misinterpretation becomes possible, (e.g. dutch **vernemen → to learn** (gain knowledge or skill // to be informed), dutch **foto → picture** (painted thing // photograph)).

B-4, which is a good equivalent, except that it has another argument structure than the word it should translate (e.g. italian **Il futuro di X, che si PREVEDE brillante** [the future of X, which itself-foresees brilliant] → dutch **voorzien** is normally a correct translation for italian **prevedere**, but it has the wrong argument structure; a paraphrase like **die zich als ... aankondigt** has the right argument structure, but the image of someone looking into the future is lost).

For reasons of this kind, candidates for the status of "adequate translation of Ss" can be ranked [3], if not with absolute scores on some quality measuring device, then at least with relative scores (i.e. among each other), reflecting how well they render some content aspect of Ss.

In table 1, Pi stands for parameter i, some content aspect of Ss; T(Ss) is the set of candidate translations of Ss. Scores in table 1 are relative: T(Ss) = {A,B,C,D,E,F,G,H,I}. For each Pi, [{A,B,C,D,E,F,G,H,I}, ranking_according_to_Pi] is a linearly ordered set, but [{A,B,C,D,E,F,G,H,I}, global_ranking_according_to_P(l-4)] is only a partially ordered set: e.g. A and B cannot be compared, as A scores better for parameter P1, but worse for parameter P4.

Without further information about interactions or relative importance of parameters, an element of T(Ss) can be considered to be globally better than another such element only if it has a better score for at least one parameter, and the same score for all other parameters.

[3] Whether "can be ranked" implies "can be ranked automatically by a computer", is another question.

	a	b	c	d	e	f	g	h	i
P1	3	7	1	5	5	3	1	8	8
P2	7	8	1	4	6	4	8	1	1
P3	3	4	2	4	4	7	1	7	7
P4	6	4	2	4	6	6	6	1	2

Table 1.

In this poset of acceptable translations, MAX = {C, G, H}. Note that if candidates g and h did not exist, C would be the maximum of the poset this table represents, such that MAX = {C}.

So MAX has three elements, and the table does not provide any means to justify the eventual choice of the human translator, as the rankings on P1 to P4 are disjoint. This means that we need more than just these parameters Pi; we also need a measurement for their relative importance.

In this respect, Juliane House's 1981 book proposes a theory I will try to synthesize as follows: a number of text aspects, 'parameters', contribute to the text's **function.** A perfect translation, according to House, would have:

- the same function as its SL counterpart,
- a match along the dimensions which are found (in the course of the analysis [4]) to contribute in a particular way to the text's function. This means that it should obtain the same function with means parallel to those the source text 'uses' to obtain that function.

Note that this second criterion reminds us of an excessively strict approach to compositionality, which implies that not only the meaning of a text is compositional, but also that translations are to be obtained in a compositional way. This is not something I can agree with.

[4] Note that House does not believe a formalization or automatization of source text analysis to be possible.

To quote House:

> "In the case that two Target Texts have mismatches on different parameters, clearly this simple quantitative comparison is inadequate. We may say, however, that the degree to which a particular parameter is marked in a Source Text, is the degree to which it contributes towards that ST's function, i.e., for an individual text, a relative hierarchy in terms of parameters is feasible. Therefore mismatches on parameter A in TT1 may be seen as contributing to a greater or lesser extent to a functional mismatch between ST and TT1 than the extent to which mismatches on parameter B in TT2 contribute to a functional mismatch between ST and TT2, given that the relative importance of parameters A and B is established by the ST analysis."

From the MT point of view, this implies that there are (at least!) three interacting subcomponents which determine how well the system performs in evaluating and selecting candidate translations.

1. A subcomponent capable of calculating **semantic distance** (= distance between target language items and the source language item of which they should duplicate the semantics); i.e. a device that gives scores for individual parameters ('disciplines') as there are in table 1. As I said already, information on semantic distance alone would leave us in a deadlock position, as parameters can be mutually incomparable.

2. A subcomponent capable of determining the **relative weight** of individual subcomponents of the source text (=the relative value of their contribution in making up the global communicative value of the source text). This is the device that helps the systems out of the deadlock; in simple words, its function is to eliminate those candidate translations that make errors in crucial areas, in favour of those that perform well in important areas.

3. Some component handling **translational movement:** e.g. compensation and split/join operations on sentences.
   Assuming that split/join operations on sentences are well-known phenomena in the world of translation, I just give some examples of what I mean with 'compensation'.

e.g.	(1)	This person	is a bachelor. (= 'male' + 'unmarried')
(Dutch)	(2)	Deze persoon	is ongehuwd (the feature 'male' is lost)
→	(3)	Deze MAN	is ongehuwd.

(Movement of the feature ‘male’ to some other constituent)

e.g. (4) Rosa Luxemburg is floating in the Landwehrkanal
(‘floating’ implies that the subject is a dead body)
(German) (5) Rosa Luxemburg schwimmt im Landwehrkanal
(= swimming OR floating)
→ (6) Die Leiche von Rosa L. schwimmt im [...]
(= the dead body of R.L.)

It is not difficult to see why a component handling a phenomenon like compensation has to interact with the two previously mentioned devices: without any information from such a device, e.g. sentence (3) would be eliminated, as it is made up of a subject-NP and predicate VP which are both bad translations of their source language counterparts. The essence of compensation is exactly that it allows good translations to be made out of seemingly incorrectly translated parts.

Although compensation and the splitting and joining of sentences seem to be common techniques for human translators, till today virtually no systematic research has been done in order to formalise these. Some interesting work has been done, however, in the somehow related area of ‘Wide-Range Restructuring’. In translation between unrelated languages (say English and Japanese) Wide Range Restructuring is needed to bridge the profound stylistic gaps between source and target language. (Cf. Tsutsumi 1990).

5 Adequacy levels for MT systems

As I said in my introduction, even if we consider an MT system to be a formal grammar producing (s, S’) couples, it is not evident to formulate MT counterparts for the classical adequacy levels (observational, descriptive, explicative) for normal monolingual generative grammars.

My idea is then that abstract adequacy levels for MT systems should immediately reflect how good some system is at evaluating candidate translations, and how well it copes with ‘untranslatable items’. This gives us the following highly tentative list of possibilities:

Level A. Granting the status of language[5] to the coverage of current operational MT systems, the first level resembles what is (almost) achieved by some good **demo** language pairs. It delivers translations, and that is all there is. More concretely,

[5] In a theoretical sense, the coverage of some grammar is always a language. What I mean here is rather something like “significant subset of a natural language”.

- The system produces **at least one** acceptable S-T translation for each translational unit a professional S-T translator [6] considers to be S-T translatable.
- In the case of multiple translations, what the system delivers is just an arbitrary subset of the set of acceptable translations, without being able to express some preference.
- In the case of untranslatable units, no translation comes out, but the system is unable to explain why it fails to find one.

I would not object if someone calls this 'observational adequacy'.

Level B. In the case of multiple translations the system produces exactly the subset **MAX** (see above), because it is able to use information on semantic distance.

In practice, this means that such a system, instead of blindly giving a unordered set of translations on the basis of (among other things) denotational 'equivalence' relations stated in its lexicon(s), it eliminates all candidate translations that are worse on all fronts (i.e. for all parameters in the table) than some other candidate, before delivering its set of translations.

This level of adequacy can be split up, in function of some 'extras':

- (B+) the above holds only within the limits of compositional translation, or is complemented with the capability to take into account the effects of translational movement.
- (B++) the system explains where the problem is in the case of untranslatable units, (e.g. 'A has a larger denotation than A', the only possible 'candidate'; or 'The Target Language has no lexical equivalent for A').

Level C. The system is able to distil a hierarchy of source text elements in function of how important it is that they are maintained in translation. In other words: such a systems knows about the relative weight of individual components of a text, and how this interacts with semantic distance and translational movement.

[6] I would like to stress the theoretical importance of the human professional translator; just as in generative grammar the ideal speaker-hearer has always the last word in translation matters. But to make up for the fact that translating is not really an innate capacity, and that the status of good translation is to a certain extent determined by tradition (norms, and schools of translation), I have substituted the mere bilingual with the "human professional translator".

Practically this means that instead of giving the MAX subset of potential translations, the systems gives exactly those elements of MAX that a professional S-T translator would call the best translation(s).

This also means that in the case of untranslatable items, the system is capable of suggesting the most appropriate approximate translation in function of which aspects are more important than others.
It is important to note that there is also a procedural interpretation of the description of these capabilities: i.e. indications of what MT developers should look for in order to produce more powerful MT systems.

References

1. Chomsky, Noam, 1957. *Syntactic Structures,* The Hague: Mouton.
2. Chomsky, Noam, 1965. *Aspects of the Theory of Syntax,* Cambridge: M.I.T Press.
3. House, Juliane, 1981. *A Model for Translation Quality Assessment,* Tübingen: Gunter Narr.
4. Tsutsumi, Taijiro, 1990. *Wide-Range Restructuring of Intermediate Representations in Machine Translation,* in "Computational Linguistics", 16/2, pp. 71-78.