Interlanguage and Set Theory

Atle Ro
Bergen

Abstract

If one is to exploit the notion of interlanguage in error diagnosis systems, a precise definition of this concept is useful. We will define interlanguage set theoretically, on two different levels. A comparison of a first language and a target language on the X^0-level makes it possible to compare structural similarity between languages, and a comparison on the V_T-level enables an explication of second language acquisition. We also comment on the limitations of set theory applied to interlanguage, and propose some augmentations which are needed in a theory of interlanguage that is to give a satisfactory account of interlanguage data.

0. Introduction

The notion 'interlanguage' alludes to a language "between" two (or more) languages, i.e. a target language (Lt) norm which a student is trying to achieve, and his first language (L1). The interlanguage has characteristics of both of these languages. The nature of the blending, or how "between" is to be interpreted, however, has always been vague in second language acquisition (SLA) literature. In this paper, we will try to make the concept so clear that it can be exploited in a computational system for diagnosing second language errors. In our study, Lt is Norwegian, and L1 is Spanish.

The main features of interlanguages which will be used in the diagnosing system, are overgeneralisation of Lt rule statements and transfer from L1. In the diagnostic system, overgeneralisation will be implemented as constraint relaxation along the lines of Douglas and Dale (1992), and transfer will be implemented by means of an alternative L1 based grammar. Transfer is understood in the sense it is used in SLA research (cf. Odlin (1989)), not in the sense of machine translation (although the planned system bears resemblances with transfer based MT systems).

In the main section of this paper 'interlanguage' is defined set theoretically. In Section two some features which we want in a theory of interlanguage, but which fall outside the set theoretical study in section one, are discussed.

Proceedings of NODALIDA 1993, pages 217-223

1. Interlanguage – a Set Theoretic Definition

We want to define foreigners' interlanguage or second language in terms of the target language they aspire to master and their first language. Let the first language grammar be G1, and L(G1) the language generated by G1. Let the interlanguage grammar be Gint, and L(Gint) the language generated by the interlanguage grammar, i.e. the interlanguage. Furthermore let the target language grammar be Gt, and L(Gt) the target language.

We then define G1 as:

$$G1 = \langle V_T, V_N, V_{X0}, \{S\}, P \rangle$$

where V_T is lexical entries of L1, V_N = {NP, S, PP, VP, N, ...}, i.e. the set of grammatical categories, and V_{X0} = {N, V, A, P, CONJ, ADV}, i.e. X^0-categories. S is the axiom, and P the grammar rules of G1. We assume that V_T and V_N are disjoint sets. V_{X0} is a proper subset of V_N. Strings over V_{X0} will be called X^0-strings. An X^0-string is e.g. *Det N V N*, whereas a terminal string (a string of terminal symbols) is e.g. *a man eats sushi*. Gt is defined in the same way.

The languages generated by two grammars can now be compared on two levels, the V_{X0}-level and the V_T-level, and 'interlanguage' understood in terms of these two grammars. Both approaches are useful. The former makes it possible to express the degree of structural similarity between languages (with the possibility to explain both positive and negative transfer), and the latter enables one to explicate processes of language acquisition. We will first consider the former approach.

1.1 Comparison on the V_{X0}-level

It is possible that different grammars can generate languages which are equal on one level, but not on another. If the two languages are equal on the V_{X0}-level, i.e. that their sets of V_{X0}-strings are equal, the possibility that the grammars which generate them are different exists, but it is probable that the grammars are quite similar. On the other hand, if the two languages are different on the V_{X0}-level, the rules of the two grammars cannot be equal. As a working hypotheses or plausible assumption we propose that an interlanguage in terms of V_{X0}-categories, is something like what we see in figure 1.1:

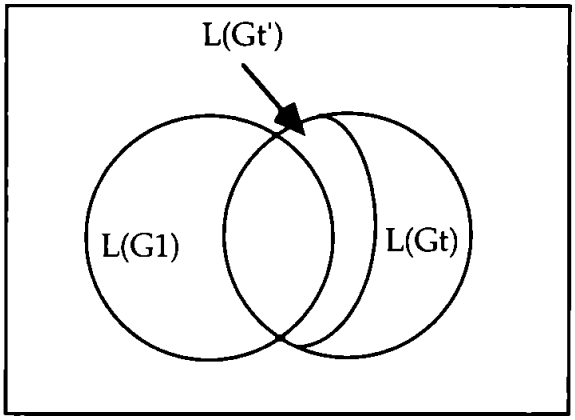

FIG 1.1: Interlanguage w.r.t. V_X0

L(G1) and L(Gt) overlap, and the degree of overlap is determined by the similarity between the two grammars. We make two assumptions about interlanguages:

1) the interlanguage user has a representation of Gt, which we will call Gt', and furthermore, Gt' is not a complete rendering of Gt. This implies that we assume that the full range of possibilities of Gt are not exploited in L(Gint), which means that L(Gt') is a subset of L(Gt).

2) L(Gint) contains strings which are not admitted by Gt, but by G1.

So preliminary we say that an interlanguage L(Gint) is the union of L(G1) and L(Gt') w.r.t V_X0.

1.2 Comparison on the V_T-level

Let us first compare G1 and Gt. Assume that both grammars have the same V_N. Assume further that the axiom is the same, and that the terminal vocabularies are disjoint. Thus the languages are completely different as in figure 1.2.1.

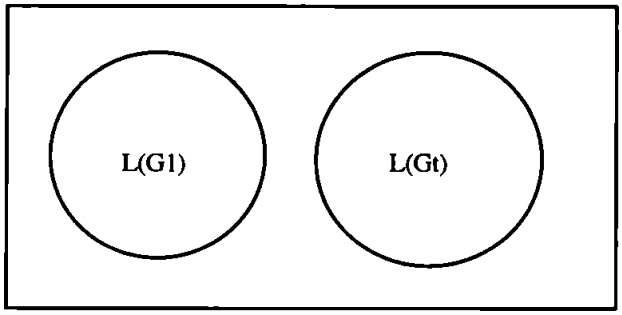

FIG 1.2.1: Comparison of L(G1) and L(Gt)

As soon as L(GI) $\cap$ L(Gt) $\neq \varnothing$ we have an interlanguage w.r.t V_T. This is, however, impossible, because the terminal vocabularies are disjoint. If we, on the other hand, assume that Gint = G1, where V_T corresponds to the empty set, this provides a model of interlanguage at the initial state. We assume that at the outset Gint is very similar to G1, at least w.r.t. production rules, but as it develops, rules from GT are added (acquired). At the outset, the vocabulary of Gint is very small, but increases during the acquisition.

We then define Gint like this:

$$\text{Gint} = \langle V_{T_t}, V_N, V_{X^0}, \{S\}, P \rangle$$

where V_{T_t} is lexical entries of the target language, V_N = {NP, S, PP, VP, N, ...}, i.e. the set of grammatical categories, V_{X^0} = {N, V, A, P, CONJ, ADV}, i.e. X^0-categories, S is the axiom, and P can contain grammar rules of both G1 and Gt.

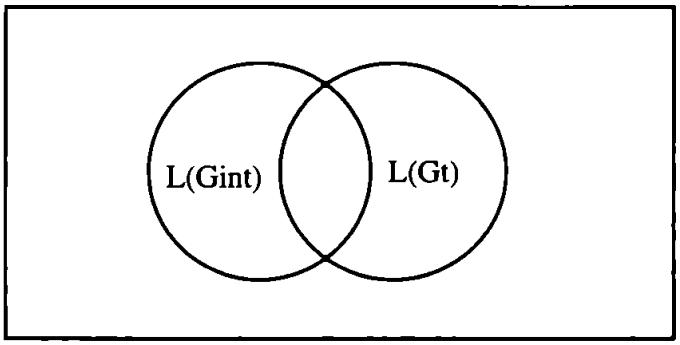

FIG 1.2.2: Interlanguage w.r.t. V_T

The intersection of the two languages is the subset of the interlanguage which is correct (cf. fig. 1.2.2) w.r.t. Gt. As the interlanguage develops, and the similarity between L(Gint) and L(Gt) increases, L(Gt) will be eclipsed to a varying degree by L(Gint).

An important theoretical issue is the following: how can we account for the fact that some rules from G1 are not present in Gint? It is hardly surprising that some rules of Gt are not present in Gint, we can explain this in terms of incomplete language acquisition. But how are some G1 rules excluded from Gint? Do rules from the Gt and G1 exclude each other in Gint, do they exist side by side, or both? The answers to these questions will tell us much about the mechanisms of SLA.

2. Interlanguage competence

Starting with the set theoretic study in section one, we already have a theory of interlanguage competence, albeit a very simple one. In this section we will first elaborate some of the assumptions made in section one, and then go on to discuss some augmentations which will bring the theory closer to the data we want to account for.

We assume that L(Gt') is a subset of L(Gt) (cf. section 1.1). How can we justify such a claim? Imagine that a rule of Gt is not in Gt', nor in G1. This rule licences a special type of strings (e.g. it-cleft sentences). Now there will be no instances of this type of strings in L(Gt'). It is natural that advanced rules of Gt are acquired at a later stage than the more basic ones like S -> NP VP, and this supports our assumption.

We also assume that L(Gint) contains X^0-strings which are admitted by G1 and not Gt. Examples of this kind of erroneous strings are Norwegian pseudo-sentences[1] displaying pro-drop and V2-violations. These errors in Norwegian are admitted by a Spanish G1.

In section 2.2 we claimed that Gint does not contain all the rules of G1. because some interlanguage errors (w.r.t. Gt) that should be accounted for by G1 rules never appear in interlanguage data. Therefore it is an oversimplification to say as we did in section 1.1 that an interlanguage is the union of L(G1) and L(Gt') w.r.t. V_{X0}. As for L(G1), we are dealing with a subset which is diminishing along with the progress of the student.

Now we will introduce some augmentations to the set theoretic account we have made so far. First we will introduce syntactic features in rules and lexical entries. We want to replace the simple rule format of section one with rules that refer to syntactic categories which are feature-bundles or attribute-value matrices (AVM's). We further assume that lexical entries in Gint may be underspecified w.r.t. syntactic features (compared with the corresponding Gt lexical entries), or even have wrong values for features. This augmentation will enable us to account for agreement

[1] By 'pseudo-sentence' is understood an ungrammatical string which is almost a sentence.

errors, non-finite verbs as heads of sentences etc. If we assume that rules refer to feature-bundles, and such rules of Lt are learnt in an incomplete or imprecise fashion, we can imagine that Gint partly is a deprecised and perhaps incomplete version of Gt. And errors like (1) can be accounted for.

(1)	*noen*	*liten*	*prosjekt*
	some-pl	*small-sg*	*project(s)*

This means that we must revise our notion of L(Gt') as a subset of L(Gt) (cf. fig. 1.1). L(Gt') contains strings which are not in L(Gt), like (1).

A theory of interlanguage competence should account for lexical transfer from L1. By lexical transfer we mean that Lt lexical items are assumed to have the same syntactic information associated with them as the corresponding L1 lexical items. With 'corresponding' we mean 'having the same meaning'. The example in (2) illustrates negative lexical transfer from Spanish.

(2)	**Jeg*	*kunne*	*ikke*	*svare*	*til*	*ham.*
	I	*could*	*not*	*answer*	*to*	*him.*

The Norwegian verb *svare* subcategorises for an object NP, while the Spanish verb with the same meaning, *responder*, subcategorises for a PP headed by the preposition *a* (to). If we assume that lexical items of L1 and Lt are linked to each other when they have the same meaning, subcategorisation information from the L1 lexical item can be used in generating the interlanguage string. Thus lexical transfer of the kind illustrated in (2) can be accounted for. To sum up, we assume that the interlanguage competence has access to the L1 lexicon, and lexical items of Gt and G1 are connected as outlined above. L1 word forms are not, however, used as "terminal vocabulary" in interlanguage strings.

3. Conclusion and future work

A set theoretical definition of the concept 'interlanguage' has been given, and a theory of interlanguage competence has been outlined. Future work will exploit the insight from this concept of interlanguage in developing a system for diagnosing ill-formed input based on overgeneralisation of Gt, and negative transfer from L1. This will be done by means of constraint relaxation of Lt rules, and an alternative L1 based grammar.

Acknowledgements

I wish to thank my supervisors Helge Dyvik and Torbjørn Nordgård for fruitful comments while writing the paper. I also wish to thank Patrizia Paggio and Ebbe Spang-Hanssen for pointing out limitations of set theoretical studies of interlanguage after reading an early version of the paper.

References

Douglas, Shona and Robert Dale. 1992. *Towards Robust PATR.* In *Proceedings of the 15th International Conference on Computational Linguistics*, Vol. 2, pp. 468–474, Nantes, France.

Odlin, Terence. 1989. *Language Transfer.* Cambridge University Press.