

An Operational Definition of the Emergence Criterion

GABRIELE PALLOTTI

University of Modena and Reggio Emilia, Italy

Although acquisition criteria are a fundamental issue for SLA research, they have not always been adequately defined or elaborated in the literature. This article critically scrutinizes one such criterion, the emergence criterion, proposing an explicit, operational definition. After discussing emergence as a theoretical construct, the article addresses several points involved in its operationalization. These points concern all stages of a research project, from data collection to data organization and analysis. A concrete example is provided, leading to the formulation of an emergence criterion for the acquisition of two grammatical structures of Italian as a second language. Issues of reliability and validity are also discussed, providing indications for future research.

ACQUISITION CRITERIA AND SLA RESEARCH

Acquisition criteria have a long history in first and second language acquisition (SLA) studies (e.g. Cazden 1968; Brown 1973; Dulay and Burt 1974; Tarone *et al.* 1976; Andersen 1978; Meisel *et al.* 1981; Vainikka and Young-Scholten 1994; Pienemann 1998; Bardovi-Harlig 2000). We need acquisition criteria if we are to be able to make replicable and falsifiable claims about the order in which different linguistic structures appear in an interlanguage (IL). For a statement like 'A is acquired before B' to be replicable and falsifiable, one needs to provide an operational definition of the construct 'acquisition'. Acquisition criteria are in fact operational definitions which allow us to determine, for a given interlanguage sample, whether a structure has been acquired or not.

As with all operational definitions, the representation of acquisition in terms of specific criteria is to some extent arbitrary. Any criterion sets a threshold dividing presence from absence, or one level from another. One may indeed formulate more complex criteria, segmenting knowledge of a linguistic structure into several degrees or levels, but the shift from one level to the next will always be stated in categorical, yes/no terms. In sum, interlanguage development is often described as a series of stages, while it is well known that many acquisition phenomena are continuous (Sharwood Smith and Truscott 2005).

However, it is also true that the interlanguage develops through a series of qualitative 'leaps', which MacLaughlin (1990) calls 'restructuring'. Representation of the linguistic system changes over time, so that today's interlanguage is qualitatively different from (and not just more fluent or automatized than) that of yesterday. In other words, some structures may appear in today's interlanguage that did not exist yesterday.

For all these reasons, acquisition criteria have been extensively employed in first and second language acquisition research. They often tend to be formulated as accuracy percentages—a structure is acquired when it is used correctly in 60 per cent (Vainikka and Young-Sholten 1994), 75 per cent (Ellis 1988), 80 per cent (Andersen 1978), or 90 per cent (Dulay and Burt 1974; Bahns 1983) of cases. This raises several issues. The first is that choice of the criterion level seems rather arbitrary and no author has provided convincing theoretical reasons for maintaining that a certain threshold is a more valid indicator of acquisition than another. Such choices, however, have relevant implications. As Hatch and Faraday (1982: 182ff) have shown, applying two different acquisition criteria—one formulated as 60 per cent accuracy, the other as 80 per cent—leads to different acquisition orders of the same structures with the same data set. Secondly, all these levels are at the high end of an accuracy scale, equating 'acquisition' with 'mastery'. The resulting developmental sequences will thus refer to the order in which linguistic structures are mastered, which may not correspond to the order in which they first entered the interlanguage (Pienemann 1998: 137).

A further concern is that accuracy with respect to L2 norms is not a valid indicator of interlanguage development. An interlanguage is in fact a linguistic system which should be described in terms of its own internal regularities, rather than by counting errors relative to L2 norms. As Sorace (1996: 386) notes, if one is interested in reconstructing the learner's grammar, 'the evaluation of the distance between native and nonnative grammars becomes an irrelevant criterion'. Describing IL development in terms of L2 accuracy thus entails what Bley-Vroman (1983) called the 'comparative fallacy'.

In order to overcome these limitations some authors have proposed acquisition criteria based on the emergence of linguistic structures (e.g. Meisel *et al.* 1981; Bahns 1983; Hammarberg 1996; Pienemann 1998; Bardovi-Harlig 2000). Three reasons are given to justify this. First because the risk of committing the comparative fallacy is reduced (Lakshmanan and Selinker 2001). In fact, by focusing on the very first uses of a new structure—rather than asking 'how much' it is supplied or 'to what extent' it is correctly used—one can identify more clearly any regular distributional patterns which may not correspond to any of the L2 rules. Secondly, emergence of a structure seems to be a more constant and less arbitrary landmark with respect to accuracy levels set anywhere between 60 and 90 per cent. Finally, emergence focuses on the order in which structures first appear, which represents a qualitative restructuring of the interlanguage.

This is considered to be an important turning point, as Pienemann (1998: 138) has argued:

The one cut-off point which remains constant... is the point of *emergence*, which is also relevant for other reasons. From a speech processing point of view, emergence can be understood as the point in time at which certain skills have, in principle, been attained or at which certain operations can, in principle, be carried out. From a descriptive viewpoint one can say that this is the beginning of an acquisition process, and focusing on the start of this process will allow the researcher to reveal more about the rest of the process.

Formulating an emergence criterion, however, does not mean simply moving the accuracy threshold to 10 or 20 per cent, for this solution would still fall foul of the comparative fallacy. Even if the criterion were applied to specific IL rules rather than to rules of the target language, several other questions would remain. For example, what does 'accuracy' really mean and how can it be measured? Do under-suppliance errors count in the same way as over-extensions or malformations? How can unanalyzed formulas be accounted for? What theoretical grounds can be invoked to demonstrate that the selected criterion is a valid indicator for the 'emergence' construct? Such questions concern all acquisition criteria used in SLA research and have not yet received satisfactory answers (Jansen 2000; Norris and Ortega 2003; Jordan 2004).

These are the kinds of problems with which this article is concerned. The exposition begins with an overview of Pienemann's approach to formulating an emergence criterion (EC). The necessary steps for defining and operationalizing emergence will then be discussed. These include theoretical definition of the construct, identification of behaviours indicating its presence or absence, specification of elicitation procedures, data organization and scoring, data analysis and interpretation, assessment of reliability and validity. Finally, a concrete example will be given of how an EC was formulated in a project on the acquisition of Italian as a second language.

Due to the empirical research upon which this article is based, only the emergence of inflectional morphology will be dealt with. Some of the issues involved concern the acquisition of syntactic constructions as well, and in principle the criterion which is argued for here could be adapted to syntax. However, a detailed treatment of this point would require a discussion that cannot be pursued here.¹

PIENEMANN'S DEFINITION OF THE EMERGENCE CRITERION

Pienemann (1998) proposes Processability Theory as a way to account for developmental sequences in interlanguage. According to the theory,

a set of processing procedures is required for the production of L2 linguistic structures. These procedures develop in the second language in an implicational order: each of them requires lower-level procedures in order to be deployed, and stages in this 'processability hierarchy' cannot be skipped. Linguistic structures will thus *emerge* in the interlanguage only when the required processing procedures have been acquired.

Emergence is defined as 'the first systematic use of a structure, so that the point in time can be located when a learner has—in principle—grasped the learning task' (Pienemann 1984: 191). In order to establish whether a structure has emerged or not, an emergence criterion must be formulated and operationalized. Pienemann (1998) discusses this issue at length, proposing the following steps for establishing the point of emergence of grammatical morphemes:

- 1 constructing distributional tables displaying form–function relationships;
- 2 testing whether these relationships are systematic, or merely the result of random hits (i.e. overgeneralizations);
- 3 testing whether the inflected words result from the productive association of a grammatical morpheme with a lexical stem or are memorized, invariant formulas.

With respect to the construction of distributional tables, Pienemann introduces the notion of 'factorization' (1998: 159) as a way to avoid the comparative fallacy. For example, a learner of a fusional language like Swedish, in which grammatical morphemes usually carry more than one grammatical meaning, may have developed a systematic opposition between \emptyset and *-a* for singular and plural adjectives, respectively. In Swedish, however, *-a* is actually used for all plural adjectives, while singular forms may end with \emptyset , *-a* or *-t* depending on gender, definiteness and on whether the adjective is attributive or predicative. By following the simple \emptyset /*-a* opposition, the learner will produce many 'wrong' forms, according to the target language rules; however, if other diacritic features such as gender and definiteness are 'factored out' of the analysis, what one sees is the emergence of a clear interlanguage rule 'use \emptyset for singular and *-a* for plural'.

Having constructed distributional tables, the next step is to ascertain whether form–function associations are systematic. Pienemann (1998: 122–9) provides a detailed discussion of a learner's acquisition of verb conjugation in German, concluding that what matters for establishing the point of emergence is not just the percentage of suppliance in obligatory contexts, but also specificity of application. A learner may go through a period of asystematic, random use of a morpheme, which cannot be identified with emergence. Thus, in a distributional table, one should not just look at percentages of application, but also at percentages of over-extension.

Finally, Pienemann warns against confusing emergence with the production of memorized chunks, and suggests the need for 'lexical

and morphological variation' (1998: 130) as a requisite for attributing emergence.

Pienemann's detailed treatment of the emergence criterion constitutes an exception and a significant improvement in the field of SLA research. The notion is theoretically well-founded and many of the methodological problems involved in its operationalization are convincingly worked through. However, the exposition proceeds through a series of illustrative examples and one is left wondering how it could be formulated in general terms and extended to other contexts.

In short, although Pienemann has gone quite a long way towards defining emergence both theoretically and operationally, 'the question of exactly what constitutes the acquisition of each level is not entirely resolved' (Jordan 2004: 227). The aim of this article is to develop the notion of emergence as a theoretical construct further and to spell out what is necessary for giving a fully explicit operational definition of the EC.

DEFINING THE CONSTRUCT

In order to operationalize a construct such as 'emergence' the first step that needs to be taken is to provide a theoretical definition, framed within the current state of research. Following Gass and Selinker (2001), SLA can be seen as a process involving the following components: (a) apperceived input; (b) comprehended input; (c) intake; (d) integration; (e) output. 'Output' can in turn be analyzed in finer detail. Initially, structures may be produced within memorized chunks of language. In the case of grammatical morphemes, they first appear in inflected words directly retrieved from memory, without the learner being able to separately analyze the form and function of lexical stem and grammatical affix. These memorized chunks play an important role in acquisition, as they provide a repertoire of forms that can be subsequently analyzed. This analysis begins with a few alternations limited to some words, gradually extending to more, and possibly all, contexts, leading to what is commonly called a 'rule'. Researchers' views differ on how they characterize this final state of productive use and the importance they attribute to memorized chunks in later stages of acquisition, but they all recognize that after a certain point learners are able to generalize beyond the examples they have been exposed to and create novel constructions (Weinert 1995; Ellis 2002, 2003).

Another common finding is that this process of productively inflecting lexical stems may begin with a phase of random variation, in which the newly discovered morphemes seem to be attached to lexical items without any clear, systematic function (Towell *et al.* 1993; Ellis 1999). Hence the acquisition of a new structure may be characterized, on the one hand, as a path from 'formula' to 'low scope pattern' to 'productive use' (Ellis 2002), on the other, as a path from 'invariant default forms' to 'non-functional allomorphic variation' to 'distributional restructuring, functional specification

and increasingly target-like use' (Housen 2002). These two paths are intertwined, and at a given time a learner may produce the same grammatical morpheme as part of unanalyzed chunks, in limited lexical alternations, with a proportion of free, asystematic variation together with some incipiently systematic application. The resulting interlanguage grammar would have a high degree of indeterminacy, and its 'rules' or 'regularities' should be stated in probabilistic terms (Klein and Dittmar 1979; Berdan 1996; Sorace 1996; Sorace and Keller 2005).

Given this overall picture of the acquisition of L2 structures, the following definition of emergence is proposed, paraphrasing and integrating Piennemann (1984):

Emergence refers to a point in time corresponding to the first systematic and productive use of a structure.

The succinctness of the definition requires a number of qualifications to be made of the terms used.

Structure

By linguistic structure we mean, in terms of inflectional morphology, a systematic association of a phonological form (e.g. in English, /s/) with a grammatical function (e.g. 'plural') (Stankiewicz 1991).

Use

Emergence is defined as first systematic *use*. This means that—even if acquisition involves receptive processes like apperception, comprehension, intake, and integration—emergence refers to the point when the structure begins to be produced. 'Production' too needs to be qualified. It is in fact recognized that some learners may produce structures based on quite different sources, on implicit and explicit knowledge. It is difficult to empirically demonstrate for every instance of language production if one or the other, or both, are involved (Hulstijn 2005), although 'this is a question of considerable importance to SLA researchers, for it is implicit rather than explicit knowledge that is deemed indicative of whether acquisition has taken place' (Ellis 2001b: 253). In the present definition, emergence is understood as referring to unplanned and unmonitored use of the structure, which most likely involves implicit knowledge.

Systematic

The very first appearances of a grammatical morpheme may be the result of random experimentation with the affix, without any clear functional meaning. It is only at a later stage that use becomes systematic, and this is a crucial requirement of the EC as defined here. In order to demonstrate

systematicity, a certain number of tokens need to be collected showing the regular association of a phonological form with a grammatical function.

Productive

In the present definition of emergence, use must be productive. With respect to inflectional morphology, this means that a given affix must be applied to a variety of lexical items including novel ones and, in principle, non-existent and thus unheard-of ones such as *wug*. It is not easy to practically discriminate between uses based on the application of productive rules from uses based on memory retrieval. The present definition of emergence however requires that *at least some* uses of the target morpheme are productive, that is that *not all* linguistic productions in which that morpheme appears are unanalyzed chunks. Such chunks, in which the phonological exponent of the grammatical morpheme is used without any apparently functional variation, are certainly important precursors of morphological emergence and they may indeed signal a stage preceding emergence proper. This stage is similar to the transitional period that Kilani-Schoch and Dressler (2002) place between ‘premorphology’ and ‘protomorphology’ in L1 acquisition. According to Kilani-Schoch and Dressler, morphology proper can be said to emerge only when ‘mini-paradigms’ appear in the data, that is when distinct inflectional forms are produced in different contexts.

First

The word ‘first’ is what distinguishes the EC from other criteria, which focus on later stages of the acquisition process. However, the association of ‘first’ with ‘systematic’ is somewhat paradoxical. When we say first, we refer to an incipient phenomenon, to a process that is just beginning, and we thus expect very few observable cases. By saying that a phenomenon is systematic, on the other hand, it is necessary to observe at least a certain number of cases—ideally, the greater the number of observations, the more reliable the conclusion of systematicity. ‘First systematic use’ thus expresses a moment in interlanguage development in which there are signs of regular, constant use of the structure, but these are the ‘first’ such signs.

Strictly speaking, ‘first’ should only apply to longitudinal data. However, one may extend the concept of emergence to cross-sectional designs as well, to mean a certain state of interlanguage in which there is at least minimal evidence for systematic and productive use of a structure—whether this level has, chronologically, just been attained or has characterized the interlanguage for a long time cannot of course be established with such data. In this sense, ‘emergence’ means ‘presence’, that is the structure has emerged (at some time) and is now being used systematically and productively.

Point in time

It follows from this discussion that emergence, as defined here, should be seen as a *terminus post quem*, that is a point in interlanguage development *after which* one can state that a structure is used productively and systematically. Emergence is thus contrasted only with what precedes it—absence or unsystematic/unproductive use—but not with what follows it. In other words, the definition specifies a lower threshold separating non-emergence from emergence, but no upper threshold, separating emergence from, say, consolidation or mastery. This way emergence becomes a resultative concept, the minimum amount of evidence necessary to state that a structure is there. When a learner begins to consolidate and automatize that structure, and even when later on she masters it with native-like proficiency, one will still say that the structure ‘has emerged’.

This separation of non-emergence from emergence is to some extent arbitrary, since interlanguage development is a gradual process. Normally there is no one specific moment in which the structure appears, as if a sudden illumination had struck the learner. However, this imposition is justifiable because dichotomous criteria are necessary if we are to be able to make any generalizable and replicable claim concerning *sequences* of acquisition—in order to state that structure A emerges before structure B, a precise criterion for establishing what ‘emerge’ means must be provided.

It may also be the case that a structure that is seen as having emerged at a certain point may not be in a later data sample. This may be due to the high variability and frequent backslidings found in interlanguage development, which is all but steadily linear. It may also be due to the fact that the data sample may not provide *enough evidence* to be able to state that the structure is present *in that sample*. This does not necessarily mean that the structure is not there in the learner’s interlanguage.

OPERATIONALIZING THE CONSTRUCT

Having defined the construct from a theoretical point of view, an operational definition must be provided in order to apply it to empirical data. In what follows, the steps involved in formulating an operational definition of emergence will be examined, from data collection to their organization and interpretation. Appendix A (available online to subscribers at <http://applij.oxfordjournals.org/>) summarizes the discussion in the form of a checklist in which methodological issues are discussed, with possible threats to validity and the features that need to be operationalized.

Data collection

EC formulation impacts in several ways on data collection. First, it specifies what kind of data will be relevant. In our definition, emergence is construed

as first *use* of a structure, which means that production data are at issue. Other formulations, defining emergence as first apperception or comprehension of a structure, will require different types of data such as repetition tasks, reaction time experiments or acceptability judgements.

Secondly, the EC operationalization specifies the minimum amount of evidence needed for meaningfully applying the criterion, that is the quantity of information in the data necessary for reaching reliable conclusions. An EC based on production data will have to specify the minimal number of contexts for using the structure, as in classical obligatory occasion analysis.

The criterion itself does not prescribe how to collect data but only what type of information must be present. If the construct is formulated in terms of language production based on implicit knowledge, then one must ensure that speech samples are produced under conditions limiting the recourse to explicit knowledge, for example with strong communicative demands. That said, any communicative task may be adequate, provided it yields a good 'data density' (Pienemann 1998), that is a certain number of relevant contexts for using the structure.² A good data density can be achieved with careful piloting of communicative tasks and, in general, by collecting rather lengthy interlanguage samples. Furthermore, if the study concerns the order of emergence of different structures, it is important that they all have comparable data density, otherwise early or late emergence of some of them might be an artifact of data collection procedures.

Data organization

Once data have been collected and transcribed, they must be organized in ways that permit subsequent analysis.

The first choice that needs to be made is whether the analysis should be based on types or tokens. In the case of inflectional morphology, analysis by type means the association of a grammatical morpheme with a single lemma, for example *book-s*. Tokens are all the repetitions of this kind of 'inflectional type' (Kilani-Schoch and Dressler 2002: 57). If a learner repeats *books* twice in a data sample, this would count as one type and two tokens. Pica (1984) has shown that type or token counts lead to placing the same learner below or above a certain criterion level, or to one learner's being ranked higher or lower in comparison with another. Hence, two versions of an acquisition criterion, one based on types and the other on tokens, may lead to different acquisition orders. In the present formulation, types are counted for assessing the degree of productivity while tokens are used in the quantitative distributional analysis, as they provide a more precise representation of actual morpheme production.

A second aspect concerns whether it is appropriate to include *all* items in distributional tables, or whether to filter some out for precise reasons. It is common practice not to include immediate repetitions of words and phrases uttered by the interviewer. But what about the repetition of words

pronounced two, five, or thirty turns previously? A decision needs to be taken and to be made explicit. The same holds for formulaic expressions. It is not easy to provide explicit and operational criteria for establishing what is formulaic: 'the issue is by no means clear cut, and chunk identification retains an irreducible intuitive dimension' (Myles *et al.* 1999: 50). This however seriously compromises reliability. Rather than relying on case-by-case intuition, authors should provide either an exhaustive list or general principles for inclusion/exclusion.

Another crucial issue here is establishing what structures the acquisition criterion should be applied to. One option is to take the structures as they are defined by the target-language grammar, which, however, falls foul of the comparative fallacy. Another is to try to specify interlanguage rules, for example by following Piennemann's suggestion of factorizing diacritic features.

Related to this is the decision as to which level of structure aggregation the EC is to be applied. Take, for example, verb conjugation: should one build a single distributional table with all the forms used by the learner and all their possible functions (i.e. the whole paradigm)? Or should separate tables be made for 1.sg, 2.sg, 3.sg, 1.pl and so forth; or one for singular and another for plural persons? The problem is not so much practical (it is easy to divide a complex table into a series of separate lines), but theoretical: are we going to apply the EC to the whole table, to single lines, or to groups of lines? What will be said to emerge: verb conjugation in the present tense, singular persons' conjugation, or first person singular conjugation? Analysis should always begin by looking at any systematic one form–one function relationship, that is from individual cells. Further examination may then reveal the emergence of more complex patterns involving several cells at a time. In order to increase reliability and replicability, it is desirable that not only the transcripts but also the distributional tables used for analysis be available to other researchers, showing what items were excluded, how data were aggregated, and how types/tokens were scored (Polio and Gass 1997).

Data interpretation

Having constructed distributional tables, we need to ask what type of pattern in the data represents the *first systematic and productive use* of a given structure.

In order to answer this question, three threats to validity must be addressed:

- 1 data are not robust enough;
- 2 use may not be productive, but formulaic;
- 3 use may not be systematic, but random.

Data robustness

What is the minimum amount of evidence needed to say that a structure has emerged? Some researchers have looked at the appearance of the very

first token (Hammarberg 1996; Glahn *et al.* 2001), but most consider valid only those results that are based on several obligatory contexts: three for Dulay and Burt (1974) and Zhang (2005); four for Meisel *et al.* (1981), Pienemann (1998: 145), Zhang (2004); five for Andersen (1978) and Pienemann (1998: 124). Four is the minimal level required by the present operationalization of the EC.

Productive use

Three complementary sources of evidence, all considered by the EC proposed here, can be invoked for claiming that *at least some uses* of a grammatical morpheme are productive. One is the presence of morphological minimal pairs: for example, finding in an interlanguage sample both *dog* and *dogs*, *year* and *years*, in singular and plural contexts respectively, would be evidence that a 'N+s' rule for plural formation is operating. A second source of evidence for the productive nature of grammatical forms comes from creative constructions—overgeneralizations like *two childs* and *I singed*, or L2 morphemes applied to interlanguage 'invented' lexemes (as in *two wakers* for 'two alarm-clocks'). A third source of evidence may be a certain amount of lexical variety in the application of a grammatical morpheme: if a sufficiently high number of words with different suffixes—say, singular and plural—were found appropriately used in singular and plural contexts, the likelihood that all of them be formulas would be lower.

Systematic distribution

If a sufficient number of tokens of the target structure are found and if they appear in such a variety of contexts as to suggest that application is indeed productive, one final problem needs to be addressed. The learner may in fact be supplying the morpheme randomly, with no clear function, in free allomorphic variation. In this case, one would not say that a systematic form–function association has emerged, but rather that the learner is still experimenting with a phonological form. A criterion must specify a way of differentiating such cases from systematic uses. What is required is not simply a threshold level of suppliance in obligatory contexts (SOC) because the learner may be using the form in 80 or even 100 per cent of the obligatory contexts while at the same time over-extending it in similar proportions in all other contexts. A measure like 'target-like use' (TLU; Pica 1984) might be more sensitive since it accounts for both omissions and overextensions. However, the TLU score does not differentiate between omission and overuse, which mean quite different things in acquisition terms. Omissions may indicate systematic but sporadic use, perhaps due to limited automatization whereas overextensions indicate a stage of unsystematic experimentation with a language *form* not clearly associated with some precise *function*.

What needs to be made clear is that a definition of the EC should pay little or no attention to the percentage of correct applications in obligatory contexts—a 5 per cent rate may be evidence for emergence while an 80 per cent rate may not. This percentage denotes consistency, automatization and spread of a structure in the interlanguage system. The only relevant indicator for stating that a structure has emerged is its specificity, targetedness, and selectivity. Only when a form begins to be used with a specific, selective function can one conclude that a rule has emerged.

This point is also made by Pienemann (1998: 126), who, however, provides no explicit indication about how to tell cases of random application from incipiently systematic ones. One answer to this problem might be a statistical test of independence such as Chi-square or Fisher's exact test, both used in the present formulation of the EC.³

Reliability and validity

'FFI [form-focused instruction] researchers rarely bother to demonstrate that their tests have validity and reliability' (Ellis 2001a: 34), and this probably holds for most SLA research (Norris and Ortega 2003: 745).

As regards reliability, interlanguage research should not be particularly concerned with its aspect of internal consistency. While this may be a useful requirement for standardized tests, it is not so relevant, and can even be misleading, if the aim is to collect data for interlanguage analysis (Swain 1990; Schils *et al.* 1991). One type of reliability check that is indeed crucial is interrater reliability. An operational definition must be clear and explicit enough to allow other researchers to reach similar conclusions on the same data sample.

However, 'making definitions clearer improves reliability but is separate from the matter of what is the *right* definition' (Scholfield 1995: 210). This raises the issue of validity. With respect to the EC, a validation argument for demonstrating that a given linguistic behaviour may be interpreted as emergence should consider how such a conclusion may *not* be valid. In the previous section several causes of invalidity were examined, pointing to the risk of mistaking emergence for some previous state of unsystematic, non-productive use, that is a 'false positive', or a type 1 error. However, if we react against such a risk by adopting very strict criteria, requiring plenty of evidence for systematic and productive use, false negatives (type 2 errors) may be produced and several learners for whom the structure *has* actually emerged would be misclassified as non-acquirers.

To my knowledge, no validation study has yet been done for acquisition criteria employed in the SLA literature. Here I will tentatively indicate a few ways in which such studies might be conducted, based on indications taken mainly from the language testing literature (e.g. Bachman 1990, 2004; Scholfield 1995; for applications to SLA see Chapelle 1998, Chaudron 2003).

Validation studies are fundamentally based on the ‘triangulation’ of various methods. The fact that a structure has emerged could thus be demonstrated on the basis of several elicitation procedures, which can be similar (e.g. a range of communicative tasks) or different (e.g. ‘wug tests’, acceptability/grammaticality judgements, elicited imitation, grammar exercises, comprehension tasks). A further source of information—which can hardly be called a method but which has been so far the most widely used—is the opinion of other researchers, that is whether they would agree on the conclusions, how valid they consider the construct’s operational definition to be, if they have used it themselves in previous research.

Another approach to validation would be to set up experiments with ‘emergence’ as the dependent variable. For instance, given the ample evidence accumulated about the positive effects of form-focused instruction on the acquisition of a variety of structures (Ellis 2001a), one might hypothesize that a given instructional treatment would make learners move from lack of a structure to its emergence. The validity of the EC would be strengthened if it were shown to consistently differentiate learners before and after the treatment.

This leads to so-called ‘predictive validity’. If we posit certain theoretical models to be valid, for example Processability Theory, or simply a well-observed developmental sequence such as that obtained in the morpheme-order studies, then we would expect that emergence of a structure predicts later or simultaneous emergence of other structures in accordance with established theories. An EC would be validated to the extent that learners classified by it as ‘acquirers’ subsequently moved to the predicted stage, while those classified as ‘non-acquirers’ did not.

AN EXAMPLE

Having discussed the steps that need to be taken to formulate an EC, a specific example of how the EC can be fully operationalized will now be outlined. Exemplification will be based on two structures of Italian, noun inflection and third person singular (3.sg) verb conjugation in the present tense. The same procedure can in principle be followed for other structures and languages with similar inflectional paradigms.

Nouns in Italian belong to two genders, masculine and feminine. Most masculine nouns take the *-o* ending for singular and *-i* for plural (e.g. *ragazzo/ragazzi* ‘boy/boys’). Most feminine nouns end in *-a* for singular and *-e* for plural (e.g. *ragazza/ragazze* ‘girl/girls’). A third inflectional class, accounting for about 20 per cent of the nouns, masculine and feminine, takes *-e* for singular and *-i* for plural (e.g. *insegnante/insegnanti* ‘teacher/teachers’). The learner must thus learn that both *-i* and *-e* express plural, and, for each lexical item, whether it requires one suffix or the other.

Italian finite indicative verbs are conjugated for person and number. The same suffixes are used for verbs belonging to different inflectional classes

(conjugations), with the exception of 3.sg, which ends in *-a* for the first inflectional class and in *-e* for the others. Here we will only be concerned with the acquisition of *-a/-e* as markers of 3.sg present tense.

Data collection

The participants in the study were six adult learners of Italian as a second language, with different L1s, who had arrived in Italy a few weeks before data collection began, having very little or no knowledge of the language. They were enrolled in a course of Italian as a second language in an Italian university; three of them were university students. The course had a communicative orientation, but some explicit grammar instruction was also provided; students attended 6 hours a week. The participants were interviewed over periods ranging from two to five months, at 2-weeks intervals.

The study was aimed at testing Processability Theory, and interviews were designed to elicit as many structures as possible that could provide evidence for the emergence of the procedures of the processability hierarchy. Besides some spontaneous conversation, the learners were asked to perform a variety of communicative tasks, including film and picture story retellings, picture descriptions, spot-the-differences, and giving instructions. Interviews lasted on average 30–45 minutes, ensuring a high number of contexts for producing the test structures. Also included was a seventh participant (Peter), whose data come from another longitudinal project on L2 Italian coordinated by the University of Pavia (Andorno 2001).

Details on the data collection procedures and elicitation materials used can be found at www.gabrielepallotti.it.

Data organization

Data were transcribed using a modified version of the CHAT-CA format. Every transcription was reviewed twice by different researchers. Distributional tables were constructed to plot the relationships between phonological endings and grammatical functions.

The following **exclusion criteria** were formulated. All nouns and verbs were reported in the distributional table, and those not considered for subsequent quantitative analyses were crossed out.⁴ These include: (a) echoes of words uttered by the interviewer in the preceding 80 words (produced by either interviewer or subject); (b) words with inaudible morphological ending; (c) uninflectable nouns; (d) immediate verbatim self-repetitions of the same word or phrase. Formulaic chunks were identified using an exhaustive list (reported in Appendix B, available online to subscribers at <http://applied.oxfordjournals.org/>). If a noun appears overwhelmingly inflected for plural in the input (e.g. *capelli* 'head hair' in Italian), it was excluded from quantitative analysis. 'Overwhelmingly' was

operationally defined as appearance of the plural form three or more times more frequently than the singular in a corpus of spoken Italian (De Mauro *et al.* 1993). Copula *be* and auxiliaries were not included in the analysis of verbs, as they belong to a closed-class set and their conjugation in Italian is highly idiosyncratic.

Turning to **factorization**, only number assignment on nouns was considered, disregarding gender. Hence, both plural suffixes *-i* and *-e* were seen as correct allomorphs for the plural morpheme, regardless of the noun's inflectional class. Similarly, both *-a* and *-e* were considered correct markers of 3.sg, regardless of the verb's inflectional class.

Both **types and tokens** were taken into consideration and analyzed. For each cell and each column 'correct' and 'incorrect' uses were calculated, based on the factorized rules described above.

All transcripts and distributional tables, showing what items were excluded/included and how they were aggregated and scored, are available at the website www.gabrielepallotti.it.

Data interpretation

Productive use

In order to demonstrate that use is productive, at least two morphological minimal pairs had to be present, that is two nouns produced in both singular and plural form (e.g. *bambino/bambini* 'child-SG/child-PL') or two verbs inflected for both 3.sg and another person/number (e.g. *mangia/mangiamo* 'eat-3SG/eat-1PL'). Each of these pairs could be substituted by a creative construction or by three pairs of correctly inflected lexemes. For example, *computri* 'computer-PL' would be counted as a creative construction, as the Italian word *computer* is not inflected for number; similarly, a verb form like *matrimonia* represents a creative construction, as the suffix *-a* for 3.sg is applied to the Italian noun 'marriage' to create the verb 'to get married'. 'Three pairs of correctly inflected lexemes' means three nouns inflected for singular and three for plural, or three verbs inflected for 3.sg and three inflected for other persons/numbers.

Systematic distribution

One of the threats to validity is that 'correct' form-function associations may be produced by the learner randomly using a form with no clear association to a given function. A statistical test should thus be applied to evaluate the independency of use of a phonological form from a specific functional context. The present EC uses Chi-square to test independence; Fisher exact test is applied in tables with expected frequencies smaller than 5. The criterion level is set at $\alpha \leq 0.05$.

Reliability and validity

Two researchers independently rated the transcripts from 15 interviews, randomly chosen from the entire corpus. For both plural noun inflection and 3.sg verb conjugation, raters agreed 13 out of 15 times on the emergence or non-emergence of the structure ($\kappa = 0.700$). Causes for disagreements were identified and resolved for subsequent analyses.

Turning to validity, it was not possible to conduct a validation study proper. However, the choices made in the operationalization of the EC can be supported in terms of logical argument and degree of coherence with previous research. More specifically, the productivity aspect of the emergence construct was addressed by the requirement of morphological minimal pairs. These have been suggested by Pienemann (1998: 127, where he speaks of 'verbs which vary morphologically') and are also consistent with the notion of 'mini-paradigms' in L1 acquisition (Kilani-Schoch and Dressler 2002). Creative constructions, whose importance for establishing productivity has been underscored in L1 acquisition since Cazden (1968), were considered to be an equally valid source of evidence. More questionable might be the choice of three 'unmatched pairs' as a substitute for a morphological minimal pair. The rationale is that it would be unlikely that all these correctly inflected morphemes were retrieved from memory as uninflected chunks, and that at least some may be the result of productive affixation. The EC proposed by Zhang (2004) also requires that a morpheme appear on lexically varied material, although with much smaller figures (two different lexical contexts are enough).

As regards systematicity, Chi-square was applied in previous research on the acquisition of inflectional morphology to test for independence of form-function relationships in distributional tables (Schumann 1987; Housen 1993).

Comparing criteria

It may be useful at this point to compare the present criterion with two other emergence criteria. The first is proposed by Zhang (2004: 451): 'the adjective suffix -de (ADJ) was viewed as having emerged if there were a minimum of four tokens of it in a sample set. In addition, the context in which -de (ADJ) occurred had to vary lexically in at least two of the four tokens'. The second criterion, by Di Biase and Kawaguchi (2002: 290), is: 'the rule is supplied more than once in lexically and structurally varied environments'.

A first difference is that these criteria are less explicit than the one proposed here. For example, little or no information is given on how the data were organized, what items were excluded from analysis, or whether interrater reliability was checked or not. Since it is not possible to access transcriptions and distributional tables, one cannot establish how the figures in the summary tables were derived. The two criteria address

Table 1: *Plural noun inflection*

	Kristen			Peter			Ahmid			Karen		
	EC	DK	Z	EC	DK	Z	EC	DK	Z	EC	DK	Z
Int 1	+	+	+	-	-	-	+	+	+	(+)	-	+
Int 2	+	+	+	(+)	+	+	+	+	+	(+)	-	+
Int 3	+	+	+	-	+	+	+	+	+	+	+	+

Note: EC = proposed Emergence Criterion. DK = Di Biase and Kawaguchi (2002); Z = Zhang (2004).

Table 2: *Third person singular present tense verb conjugation*

	Kristen			Peter			Ahmid			Karen		
	EC	DK	Z	EC	DK	Z	EC	DK	Z	EC	DK	Z
Int 1	-	+	-	/	/	/	(+)	+	+	-	-	+
Int 2	(+)	+	+	+	+	+	(+)	+	+	/	/	/
Int 3	+	+	+	/	/	/	+	+	+	+	+	+

Note: EC = proposed Emergence Criterion. DK = Di Biase and Kawaguchi (2002); Z = Zhang (2004).

the issue of productivity, although differently from the present proposal. In fact, Zhang (2004) requires that a morpheme be applied on at least two different lexemes, while Di Biase and Kawaguchi's (2002) discussion seems to imply that at least one lexeme be produced with two different grammatical morphemes (e.g. singular and plural) and at least one morpheme appear on two different lexemes.⁵ Neither work mentions creative constructions as possible evidence for productivity. Finally, these criteria do not state the possibility of random hits due to over-extensions, nor do they include a measure to quantify this validity threat.

When these different formulations of the EC are applied to actual data, they lead to the conclusions displayed in Tables 1 and 2. Three interviews from four learners were analysed using the EC proposed here and those by Zhang (2004) and Di Biase and Kawaguchi (2002); analysis was carried out on tokens. A dash (/) represents a number of contexts smaller than four. The symbol (+), used only by the present author, indicates that most but not all of the criterion's requirements are met. For productivity, this means there is only one morphological minimal pair, or one creative construction, or 3–5 different lexical items with the relevant morpheme; for systematicity, it means that the *p* value is between 0.05 and 0.10. Appendix B

(available online to subscribers at <http://aplioj.oxfordjournals.org/>) reports distributional tables for all the interviews, indicating the number of applications, omissions, and over-extensions, together with evidence for productivity and the results of statistical analysis.

As the tables show, the three criteria lead to identical conclusions in most of the cells. There are, however, a few noteworthy discrepancies. With respect to plural noun inflection, Zhang's (2004) criterion seems to be the most 'generous', as it credits all learners with emergence in all interviews except for Peter, interview 1. Di Biase and Kawaguchi's (2002) criterion and the one proposed here correlate more strongly, with the exception of Peter's second and third interview, in which not all the conditions are satisfied for the present criterion while they are for the other two. Karen's first interviews present weak evidence for emergence, which can be interpreted as '(+)' according to the present criterion or '-' following Di Biase and Kawaguchi (2002).

A similar picture emerges for verb conjugation. Here too Zhang's (2004) criterion attributes emergence to all learners at all times (except Kristen interview 1), provided there are sufficient contexts. Di Biase and Kawaguchi's (2002) and the present criterion deny emergence in Karen's first interview, as there is no morphologically varying lexeme. In Kristen's first interview, Di Biase and Kawaguchi's (2002) criterion is satisfied, as there are two morphemes with the same lexeme and one lexeme with two different morphemes. However, the number of tokens is smaller than 4, hence the '-' for Zhang (2004) and the proposed EC. In Ahmad's first and second interview both Di Biase and Kawaguchi (2002) and Zhang (2004) find enough evidence for emergence, as there is a sufficient number of contexts and the affix is applied on lexically varied material. However, the number of over-extensions is considerable ($p=0.084$ and 0.086), hence the '(+)' resulting from the present formulation of the EC.

These results can be interpreted in at least three ways.

- 1 The three criteria are all valid operationalizations of the 'emergence' construct, varying in their degree of exposure to type 1 or type 2 errors. With Di Biase and Kawaguchi's (2002) and the present criterion, the risk of false positives would be lower, but with a higher risk of false negatives; Zhang's (2004) criterion would run the opposite risk.
- 2 One criterion is valid while the others are not. For example, Di Biase and Kawaguchi (2002) and the present criterion indicate emergence, while Zhang (2004) is subject to systematic error due to formulas or random hits; or conversely, Zhang (2004) represents emergence, while the other two are systematically skewed towards the diagnosis of some later stage, such as consolidation or mastery.
- 3 The different operational definitions refer to different constructs altogether: Zhang's (2004) criterion represents, say, 'incipient emergence'

while the other two indicate something like 'mature emergence' or 'consolidation'.

The latter solution is problematic, as it implies that there are virtually as many constructs as there are operational definitions. While it is true that operational definitions are an essential way of defining constructs, a parsimony principle would suggest a limit to construct proliferation. Perhaps, at least with the criteria discussed here, the most reasonable account might be that the three operational definitions all try to capture the same construct, although they differ with respect to their degree of explicitness and the amount of evidence required for rejecting the null hypothesis. A symbol like (+) is indicative of this situation. It does not represent an intermediate stage of acquisition (somewhere in between absence and emergence), but a lower level of confidence with respect to the conclusion that a structure has emerged.

CONCLUSIONS

This article has discussed issues involved in the operational definition of an acquisition criterion based on emergence, providing an example of what might be seen as good practice, that is a criterion which is explicit, reliable, and grounded on current views of SLA. The criterion's validity has been argued for in terms of coherence with the construct's definition and previous research.

However, comparing different definitions advanced by other authors raises a problem that concerns the field of SLA as a whole, that of providing explicit, valid, and shared definitions of its key constructs. As Jordan (2004: 259) notes, with reference to another controversial notion such as language competence, 'in the history of science there are many examples of theories that started off without any adequate description of what is being explained, although sooner or later, this limitation must be addressed'. Explicating and defining constructs such as 'competence', 'emergence', or 'acquisition', and devising valid assessment procedures for them, requires a cooperative effort from many researchers working in SLA, who should also draw on the expertise of language testers. Shohamy (2000) lists in fact 'defining the construct of language ability' as the first area where language testing can contribute to SLA, while Norris and Ortega (2003: 749) conclude their review on 'Defining and measuring SLA' by stating that 'validity generalization... should also be a priority for measurement used within SLA research and should constitute the site of true collaboration between language testers or measurement specialists and measurement-informed SLA researchers'.

The aim of the present article has been to suggest ways in which the first steps of this process can be taken, by giving a more explicit and carefully formulated operational definition of a construct such as 'emergence'.

More research will be needed to validate definitions within current theories of SLA and to reach some agreement among researchers as to what should be counted as 'emergence', 'acquisition', and other equally crucial notions.

Final version received July 2006

ACKNOWLEDGEMENT

I wish to thank Bruno Di Biase, Alex Housen, Jan Hulstijn, Manfred Pienemann, Richard Young, Yanyin Zhang and four anonymous reviewers for their comments on previous drafts of this article. They are not responsible for remaining errors and shortcomings.

NOTES

- 1 The topic of this article is also related to recent debates on emergentism in SLA (see the *Applied Linguistics* 2006 special issue). In fact, advocates of emergentism need to operationally define what emergence is for their claims to be falsifiable, which makes the methodological discussion offered here relevant for such a research program.
- 2 Chaudron (2003) and Purpura (2004) provide useful reviews of tasks and procedures for assessing grammatical development.
- 3 An alternative, or a supplement, to a test of independence might be a measure of strength of association like the ratio of proportions (also called relative risk; Agresti and Finlay 1997). The EC might include a minimum threshold for this ratio, e.g. the proportion of correct applications should be 2, 3, or N times higher than the proportion of over-extensions. The present criterion does not make use of this further specification.
- 4 Notice that this is different from not including such items in the table at all. A crossed out item is still accessible, which allows the researcher or others to revise the decision or to consider the item for other types of analysis.
- 5 See also Pienemann 2007: 147: 'the researcher needs to check lexical and morphological variation (i.e. same morpheme on different words and same word with different morphemes)'.

REFERENCES

- Agresti, A.** and **B. Finlay.** 1997. *Statistical Methods for the Social Sciences*. Upper Saddle River, NJ: Prentice Hall.
- Andersen, R.** 1978. 'An implicational model for second language research,' *Language Learning* 28: 221–82.
- Andorno, C.** (ed.) 2001. *Banca dati di italiano L2*. Pavia: University of Pavia, Linguistics Department.
- Bachman, L.** 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Bachman, L.** 2004. *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.
- Bahns, J.** 1983. 'On acquisitional criteria,' *IRAL* 21: 57–67.

- Bardovi-Harlig, K.** 2000. *Tense and Aspect in Second Language Acquisition: Form, Meaning, and Use*. Oxford: Blackwell.
- Berdan, R.** 1996. 'Disentangling language acquisition from language variation' in R. Bayley and D. Preston (eds): *Second Language Acquisition and Linguistic Variation*. Amsterdam: Benjamins.
- Bley-Vroman, R.** 1983. 'The comparative fallacy in interlanguage studies: The case of systematicity,' *Language Learning* 33: 1–17.
- Brown, R.** 1973. *A First language: The Early Stages*. Cambridge, MA: Harvard University Press.
- Cazden, C. B.** 1968. 'The acquisition of noun and verb inflections,' *Child Development* 39: 433–48.
- Chapelle, C.** 1998. 'Construct definition and validity inquiry in SLA research' in Bachman, L. and A. Cohen (eds): *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.
- Chaudron, C.** 2003. 'Data collection in SLA research' in C. J. Doughty and M. H. Long (eds): *The Handbook of Second Language Acquisition*. Oxford: Blackwell.
- De Mauro, T. et al.** (1993). *Lessico di frequenza dell'italiano parlato*. Milan: Etaslibri.
- Di Biase, B. and S. Kawaguchi.** 2002. 'Exploring the typological plausibility of Processability Theory: Language development in Italian second language and Japanese second language,' *Second Language Research* 18/3: 272–300.
- Doughty, C. J. and M. H. Long.** 2003. *The Handbook of Second Language Acquisition*. Oxford: Blackwell.
- Dulay, H. and M. Burt.** 1974. 'Natural sequences in child second language acquisition,' *Language Learning* 24: 37–53.
- Ellis, N.** 2002. 'Frequency effects in language acquisition,' *Studies in Second Language Acquisition* 24: 143–88.
- Ellis, N.** 2003. 'Constructions, chunking, and connectionism: The emergence of second language structure' in C. J. Doughty and M. H. Long (eds): *The Handbook of Second Language Acquisition*. Oxford: Blackwell.
- Ellis, R.** 1988. 'The effects of the linguistic environment on the second language acquisition of grammatical rules,' *Applied Linguistics* 9: 257–74.
- Ellis, R.** 1999. 'Item versus system learning: Explaining free variation,' *Applied Linguistics* 20: 460–80.
- Ellis, R.** 2001a. 'Investigating form-focused instruction' in R. Ellis (ed.): *Form-Focused Instruction and Second Language Learning*. *Language Learning* 51, Supplement 1.
- Ellis, R.** 2001b. 'Some thoughts on testing grammar: an SLA perspective' in C. Elder et al. (eds): *Experimenting with Uncertainty. Essays in Honor of Alan Davies*. Cambridge: Cambridge University Press.
- Gass, S. and L. Selinker.** 2001. *Second Language Acquisition. An Introductory Course*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum.
- Glahn, E. et al.** 2001. 'Processability in Scandinavian second language acquisition,' *Studies in Second Language Acquisition* 23: 389–416.
- Hammarberg, B.** 1996. 'Examining the processability hierarchy: The case of adjective agreement in L2 Swedish' in E. Kellerman, B. Weltens, and T. Bongaerts (eds): *Eurosla 6: A Selection of Papers. Toegepaste Taalwetenschap in artikelen*, 55/2: 75–88.
- Hatch, E. and H. Faraday.** 1982. *Research Design and Statistics for Applied Linguistics*. Rowley: Newbury House.
- Housen, A.** 1993. 'L2 acquisition of verb morphology—a case study' in B. Kettermann and W. Wieden (eds): *Current Issues in European Second Language Acquisition Research*. Tübingen: Narr.
- Housen, A.** 2002. 'A corpus-based study of the L2-acquisition of the English verb system' in S. Granger, J. Hung, and S. Petch-Tyson (eds): *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins.
- Hulstijn, J.** 2005. 'Theoretical and empirical issues in the study of implicit and explicit second-language learning,' *Studies in Second Language Acquisition* 27: 129–40.
- Jansen, L.** 2000. 'Second language acquisition: from theory to data,' *Second Language Research* 16: 27–43.
- Jordan, J.** 2004. *Theory Construction in Second Language Acquisition*. Amsterdam: Benjamins.
- Kilani-Schoch, M. and W. U. Dressler.** 2002. 'The emergence of inflectional paradigms in two French corpora: An illustration of general problems of pre- and protomorphology' in M. D. Voekova and W. U. Dressler (eds): *Pre- and Protomorphology*. Munchen: Lincom.
- Klein, W. and N. Dittmar.** 1979. *Developing Grammars*. Berlin: Springer.
- Lakshmanan, U. and L. Selinker.** 2001. 'Analyzing interlanguage: how do we know

- what learners know?' *Second Language Research* 17: 393–420.
- McLaughlin, B.** 1990. 'Restructuring,' *Applied Linguistics* 11: 113–28.
- Meisel, J., H. Clahsen, and M. Pienemann.** 1981. 'On determining developmental stages in natural second language acquisition,' *Studies in Second Language Acquisition* 3/2: 109–35.
- Myles, F., R. Mitchell, and J. Hooper.** 1999. 'Interrogative chunks in French L2,' *Studies in Second Language Acquisition* 21, 49–80.
- Norris, J. and L. Ortega.** 2003. 'Defining and measuring SLA' in C. J. Doughty and M. H. Long (eds): *The Handbook of Second Language Acquisition*. Oxford: Blackwell.
- Pica, T.** 1984. 'Methods of morpheme quantification: their effect on the interpretation of second language data,' *Studies in Second Language Acquisition* 6/1: 69–78.
- Pienemann, M.** 1984. 'Psychological constraints on the teachability of languages,' *Studies in Second Language Acquisition* 6: 186–214.
- Pienemann, M.** 1998. *Language Processing and Second Language Development: Processability Theory*. Amsterdam: Benjamins.
- Pienemann, M.** 2007. 'Processability theory' in B. Van Patten and J. Williams (eds): *Theories in Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum.
- Polio, C. and S. Gass.** 1997. 'Replication and reporting,' *Studies in Second Language Acquisition* 19: 499–508.
- Purpura, J.** 2004. *Assessing Grammar*. Cambridge: Cambridge University Press.
- Schils, E. D. J., M. G. M. van der Poels, and B. Weltens.** 1991. 'The reliability ritual,' *Language Testing* 8: 125–38.
- Scholfield, P.** 1995. *Quantifying Language*. Clevedon: Multilingual Matters.
- Schumann, J.** 1987. 'The expression of temporality in baslang speech,' *Studies in Second Language Acquisition* 9: 21–42.
- Sharwood Smith, M. and J. Truscott.** 2005. 'Stages or continua in second language acquisition: A MOGUL solution,' *Applied Linguistics* 26: 219–40.
- Shohamy, E.** 2000. 'The relationship between language testing and second language acquisition, revisited,' *System* 28: 541–53.
- Sorace, A.** 1996. 'The use of acceptability judgments in second language acquisition research' in W. C. Ritchie and T. K. Bhatia (eds): *Handbook of Second Language Acquisition*. San Diego: Academic Press.
- Sorace, A. and F. Keller.** 2005. 'Gradience in linguistic data,' *Lingua* 115: 1497–524.
- Stankiewicz, E.** 1991. 'The concept of structure in contemporary linguistics' in L. Waugh and S. Rudy (eds): *New Vistas in Grammar: Invariance and Variation*. Amsterdam: Benjamins.
- Swain, M.** 1990. 'Second language testing and second language acquisition: Is there a conflict with traditional psychometrics?' in J. Alatis (ed.): *Georgetown University Roundtable on Language and Linguistics 1990*. Washington, DC: Georgetown University Press.
- Tarone, E., U. Frauenfelder, and L. Selinker.** 1976. 'Systematicity/variability and stability/instability in interlanguage systems,' *Language Learning*, Special Issue: 93–134.
- Towell, R., R. Hawkins, and N. Bazergui.** 1993. 'Systematic and nonsystematic variability in advanced language learning,' *Studies in Second Language Acquisition* 15: 439–60.
- Vainikka, A. and M. Young-Scholten.** 1994. 'Direct access to X'-Theory. Evidence from Korean and Turkish adult learners of German' in T. Hoekstra and B. Schwartz (eds): *Language Acquisition Studies in Generative Grammar*. Amsterdam: Benjamins.
- Weinert, R.** 1995. 'The role of formulaic language in second language acquisition: A review,' *Applied Linguistics* 16: 180–205.
- Zhang, Y.** 2004. 'Processing constraints, categorial analysis, and the second language acquisition of the Chinese adjective suffix *-de* (ADJ),' *Language Learning* 54: 437–68.
- Zhang, Y.** 2005. 'Processing and formal instruction in the L2 acquisition of five Chinese grammatical morphemes' in M. Pienemann (ed.): *Cross-Linguistic Aspects of Processability Theory*. Amsterdam: Benjamins.