Rosemarie Tracy

Language testing in the context of migration

Abstract: Over the last decades, the development of instruments measuring linguistic abilities has become increasingly important. The tests developed on the basis of the *Common European Framework of Reference for Languages* (CEFR) are a good case in point. Originally backed by and promoting a positive European vision of multilingual, mobile citizens, and celebrating the linguistic and cultural diversity of Europe as unifying factors, more recently the CEFR has faced concerns that it is being turned into an immigration-control instrument, a role it was not intended for. This metamorphosis forms the backcloth and starting point of a more general discussion of the complexity of language testing, and areas in which the CEFR should be optimized are identified.

Résumé : Au cours des dernières décennies, l'élaboration d'instruments de mesure des compétences en langues a pris de plus en plus d'ampleur. Les tests calibrés par rapport au *Cadre européen commun de référence pour les langues* (CECR) illustrent bien cette tendance. A l'origine, ce dernier a été conçu comme un instrument qui repose sur une vision européenne positive promouvant la mobilité et le plurilinguisme des citoyens et valorisant la diversité linguistique et culturelle de l'Europe en tant que trait d'union entre les peuples. Cependant, plus récemment, certains acteurs se sont déclarés préoccupés par le fait que l'on est en train de faire du Cadre un instrument de contrôle de l'immigration, ce à quoi il n'a pas vocation. Cette question est l'occasion d'engager une discussion plus générale sur la complexité de l'évaluation des compétences en langues ; elle permettra également de définir des domaines dans lesquels le CECR devrait être amélioré.

1 Introduction

Regardless of the domain under investigation, developing and applying diagnostic tools and interpreting their results raises questions beyond validity, objectivity, reliability, and fairness concerning culture, age and gender. In view of the impact for the individuals assessed, be it with respect to medical, psychological or

Rosemarie Tracy, University of Mannheim, Germany, E-mail: rtracy@mail.uni-mannheim.de

DOI 10.1515/9783110477498-007,

pedagogical interventions, the burden of responsibility is considerable. While consequences may not be immediately life-threatening when it comes to language, they can easily lead to exclusion from social or therapeutic services, from educational opportunities, from the job market, and from citizenship. As for the *Common European Framework of Reference for Languages* (CEFR; Council of Europe 2001), awareness of its potential impact has been stressed for a long time, for example: "The ways in which the CEFR is used have implications for social cohesion, access to employment, citizenship, mobility and mutual understanding in Europe" (Goullier 2007: 18). While Coste about ten years ago welcomed this "wind of assessment" as "a good wind and a healthy one" (2007: 41), nowadays researchers and practitioners are increasingly concerned about unintended impacts (cf. Van Avermaet 2016), i. e. the "dual use" potential of testing, namely the (mis)use of research originally designed for different purposes.[1] Rather than report on research, as the overall title of this volume suggests, I wish to add to the growing concern that the CEFR, after all an influential benchmark of language proficiency, could be turned into a "migration management mechanism to limit the number of migrants" (Strik 2013: 3), and "hindering integration and leading to exclusion" (ibid.: 1; also ALTE 2016; McNamara 2011).

In what follows, I will first (section 2) underscore the non-triviality of investigating and measuring linguistic competence and identify some reasons for this. Section 3 turns to selected issues relating to the CEFR, such as the problematic status of descriptors and level distinctions. On the basis of examples taken from the German test for immigrants I will ask (somewhat provocatively) what we want immigrants to think communicating in their new language and "being integrated" is all about. Finally (section 4), I briefly suggest that it is time to reconcile the CEFR with current adult second-language research and to embrace linguistic theories which can provide a cross-linguistically useful *tertium comparationis* beyond the functional approach prevalent so far.

2 Linguistic competence: invisible, (largely) inaudible, and highly complex

While testing may be "a universal feature of social life" (McNamara 2000: 3), measurement of linguistic abilities is a particularly tough case in point. But testing by means of a standardized, reliable and objective procedure is only one way of going about assessing proficiency. Diagnostic tools include more or less con-

1 See http://www.dfg.de/pm/2016_13a.

strained observation guidelines and questionnaires (to be filled out by parents, teachers or even learners themselves), corpus studies based on written texts and spontaneous speech, screenings and elicitation techniques targeting specific structures in production and/or comprehension tasks. Due to the complexity of linguistic knowledge systems and behavior, to heterogeneous learning scenarios and to inter- and intra-rater variability, none of these approaches are without problems. Nevertheless, despite theoretical and methodological complications, language assessment is important. Within educational settings, such as foreign language teaching, we want to determine whether specific criteria or standards can be met by learners or, put differently, how good we were at teaching. It is also important to find out whether someone trained as a nurse or mechanic in one language context, is capable of acting professionally in another. Knowing that Specific Language Impairment (SLI) is one of the most frequent developmental pathologies, affecting 6 – 8 % of all children, we certainly wish to identify those at risk as early as possible. Also, faced with young immigrant second-language learners, educators increasingly recognize their need for diagnostic instruments yielding insight into the current state of a learner's system, and those financing intervention programs call for evaluation procedures informing them about the effectiveness of language-fostering programs. Beyond learning contexts, we also want to be able to differentiate between stroke patients afflicted by different kinds of aphasia in order to offer appropriate interventions.

Since learners of different ages and exposure differ in acquisition strategies and outcome, we need to distinguish at least first language learners (L1), simultaneous first language learners (2 L1/3 L1), successive child L2 and L3 learners – ignoring additional languages – , and older L2/L3 etc. learners (pre- and post-puberty). To complicate matters, in immigrant contexts we also encounter so-called *heritage* speakers, i. e. second-generation immigrants (and later generations) who are exposed to their parents' attriting or quantitatively reduced L1 resources and who therefore acquire their home language to varying degrees (Polinsky and Kagan 2007). At the same time, they may be highly proficient L2 speakers of their environment's majority language.

From a research point of view, in order to better understand similarities and differences across learner types, we would require some idea of the quality and the quantity of the input available to them, as well as information concerning age of onset and length of exposure. This means that we cannot follow the principle "one size fits all", and we certainly should not simply compare second-language learners with norms obtained from L1 learners or, even worse, with some idealized notion of what native speakers can do, without actually conducting controlled research. Note, too, that any acquisition type may be coupled with specific language impairment (SLI). In the case of young L2 learners the identi-

fication of children at risk for SLI is complicated by the fact that typically developing L2 learners may initially go through phases very similar to children with SLI. It comes as no surprise, then, that pediatricians and therapists experience problems differentiating typical and atypical L2 learners and struggle with misclassifications, i.e. under- and overdiagnosis (cf. Grimm and Schulz 2014). This mirrors concerns about the reliability of rater judgment which have been voiced by experts in adult language testing (cf. McNamara 2000; Wisniewski 2010).

When it comes to choice of adequate diagnostic instruments, conflicts of interest on the part of different stakeholders are unavoidable. Policy makers want the procedure and the training of personnel to be inexpensive; pediatricians, teachers and whoever else is directly involved in assessing large numbers of participants want the process to be fast and the interpretation straightforward. At the same time, at least therapists and teachers require more than *diagnostics lite* if they want to support and trigger the next step in the acquisition process.

Some problems faced by all approaches are due to the very complexity of natural languages. "Language", after all, is only a shorthand expression for many different knowledge systems, also called "levels of representation", such as the lexicon, phonological, morphosyntactic, semantic, pragmatic and, in the literate, orthographic levels. Crucially, none of these predominantly implicit knowledge systems are directly accessible. Unlike someone's ability to swim or to drive a car, what we say, write or how we react to what others say and write, is only an indirect and often poor reflection of what we know, i.e. of what we have internally represented. But while observation in a literal sense is never an option – with some exceptions concerning phonetic and prosodic properties –, the analysis and reconstruction of successive learner systems on the basis of substantial corpora is feasible. Given longitudinal data it is possible to reconstruct individual learner profiles, to identify intra- and inter-individual developmental patterns and to draw inferences concerning abstract levels of representation. Especially in the case of unknown languages or under-researched learner types and acquisition paths, this established fieldwork methodology is an indispensable first step. But since the collection of spontaneous speech, with subsequent transcription and annotation is theoretically and analytically highly demanding and time-consuming, it does not offer itself for cross-sectional attempts at judging the linguistic proficiency of large numbers of speakers. Another serious drawback of naturalistic data collection in unconstrained conversational settings is that structures we would like to catch may never be produced. Just waiting for passives, relative clauses, or long-distance wh-questions or other patterns to "happen" would be a highly inefficient approach. So if we want to find out whether specific structures are within the reach of individual learners or learner types, we have to create conditions favoring their production and/or elicit rele-

vant comprehension responses. Again, this is no trivial enterprise (for excellent overviews see the collection of articles in Li Wei and Moyer 2008; McDaniel, McKee and Cairns 1998; Menn and Ratner 2000). Choosing to speak and choice of speech act or repertoire are voluntary activities, and a learner's ability is easily underestimated whenever he or she remains silent. On the other hand, we may also overestimate someone's competence because he or she happens to behave appropriately. Faced with utterances they cannot process, learners may mimic someone else's example, from putting on their shoes to laughing at jokes they did not really "get". At a pragmatic level, this mimicry is a clever strategy, demonstrating cooperativity and politeness. In order to elicit very specific verbal or non-verbal reactions as dependent variables, we have to design clever experiments, and what may start as a small-scale experiment may eventually turn into a standardized, norm-referenced, validated and psychometrically supported test. Hence, as pointed out by Grimm (1994: 17; my translation), "testing is a lot simpler than observing".

3 The CEFR: "an instrument of reference, not an object of reverence" (Coste 2007: 46)

The CEFR, launched in 2001, has been very successful both with respect to its educational impact and as a standard and common "currency" (cf. McNamara 2011) against which other assessment instruments can be calibrated and hence validated. The vision that gave rise to it welcomed and embraced Europe's linguistic and cultural diversity as an asset and encouraged mobility and multilingualism in an additive sense, i. e. without a threat to a speaker's first language(s). In order for the CEFR to fulfill its function as a generalized frame of reference, a "descriptive metasystem" (North 2007: 29), it (largely) abstracts away from language-specific grammars and focuses on tasks that are functionally and pragmatically equivalent across languages. In the attempt to guarantee comparability and transparency, an impressive technical and statistical machinery has been created (ALTE 2011, 2016).

Despite the remarkable career of the CEFR, there is room for improvement. Many "can do" statements contain among their descriptors quantifying ("large", "small", "short", "limited", etc.) or qualifying expressions ("relatively simple", "elementary", "complex"). Descriptors refer to vocabulary or other features the test-taker appears to be "more" or "less familiar" with, is "more" or "less likely to encounter", or to terms and tasks which are "more or less related to everyday experience". There is also reference to what interlocutors can "easi-

ly" or "partially" understand. As has been pointed out by many, the theoretical and empirical basis for these to some extent intentionally vague classifications is missing (cf. Hulstijn 2007; Quetz 2010; Vogt 2011). Similarly, because there is no appeal to psycholinguistic theories of text comprehension (cf. Quetz 2010: 188–189), it is difficult, if not impossible, to appreciate what makes texts assigned to particular proficiency levels more or less complex than others. According to some critics, the current practice of text choice and assignment to specific levels is "completely erratic" (cf. Quetz 2010: 197).

To be sure, other diagnostic instruments on the market struggle with similar shortcomings. A prominent diagnostic questionnaire aiming at the verbal behavior of bilingual children acquiring German as a second language (SISMIK; Ulich and Mayr 2003) expects raters to decide whether specific constituents, for instance articles in noun phrases, are (a) "left out most of the time", (b) "mostly deviant", (c) "sometimes deviant", (d) "mostly correct" in children's spontaneous speech (ibid.: 8; my translation). Faced with these options, the rater finds her-/himself in a dilemma. If articles are *sometimes* deviant, they could also be *mostly* correct. Crucially, we would have to know how many obligatory contexts (i.e. noun phrases requiring an article) were produced altogether, i.e. how many opportunities to produce articles learners missed or used.

The CEFR's A–B–C levels and their subdivisions suggest a progressive scale of complexity among different, consecutive states of proficiency. "The CEFR is a concertina-like reference tool that provides categories, levels and descriptors that educational professionals can merge or sub-divide, elaborate or summarise, adopt or adapt according to the needs of their context – whilst still relating to the common hierarchical structure" (North 2007: 22). However, at times, descriptors and evaluation criteria are not formulated in a manner consistent with some developmental logic, as can be seen in the following examples from the German test for immigrants (*Deutsch-Test für Zuwanderer*; Perlmann-Balme, Plassmann and Zeidler 2009: 65; my translation). At level B1 learners can "reply spontaneously and relatively comprehensively" to follow-up questions (cf. the 2nd column of line 1b), while at A2 (column 3) they "reply briefly" (the original German "knapp" is a bit more negative; with a pragmatic focus it can also mean "curtly") and/or only "partially understandably". While it seems intuitively possible to contrast *briefly* with *relatively comprehensively*, this does not mean that the former, reduced response, is communicatively inadequate. The other A2 criterion, "partially understandably", is not matched with an explicit, more positive evaluation at B1 (for instance: "is easily understandable"/"can be understood easily", etc.). Note, too, that it is unclear what could be *spontaneous* – presumably intended as a positive indicator for level B1 – about an answer to a (follow-up) question. For A2, no more rudimentary alternative or precursor, such as "can re-

spond after pausing/hesitating ...” is mentioned. Also, for A1 in the category “Speaking” (p.65; cf. line 2a, column 4), candidates “can allude to the main contents of a photo in very few words”, for A2 (column 3), they can “label a photo’s main contents briefly and very generally”. This also raises the question of how to objectively tell “alluding in very few words” from “labeling something briefly and very generally”. For B1 (column 2), test-takers “can mention main contents but also details”. While the reason for these differences may be lack of vocabulary, we might also be dealing with poor test-taking strategies or lack of understanding of the task. The latter two, at least, could be remedied by encouraging test-takers to mention all objects they can make out/recognize in the pictures they are shown.

The authors of the German test explicitly concede that for some areas tasks are not linked via a theoretically motivated rating scale: “Since migrants have to be able to perform relatively complex activities early on, some complex achievements have been assigned to the lowest level, A1, at which these activities have to be implemented” (Perlmann-Balme et al. 2009: 27; my approximate translation). Hence the decision in favor of particular tasks is motivated by language-external criteria, not by any developmental (functional or pragmatic) progression or logic.

Even though the main focus of the CEFR is not on grammatical form and correctness, selected aspects are taken into account and are included among the descriptors. Reference to grammar crops up, for instance, when descriptors contain comments such as “generally good command over grammatical structures despite clearly noticeable L1 influence” (Perlmann-Balme et al. 2009: 65; my translation). What makes this statement interesting is the expectation that raters know enough about test takers’ L1 to arrive at such a conclusion. In some cases, for instance where the rater knows that a specific L1 has no articles or no subject-verb agreement, transfer (i.e. zero articles, no overt agreement) seems reasonable. But in order to make such a statement raters need more than information about the first languages of test-takers. In addition, they need to be aware of the transitional curriculum of L2 learners, regardless of their L1. Again, from a linguistic point of view, these are complex issues, and raters should not be tempted to draw conclusions they simply cannot reach with any degree of confidence.

Since testing is all about comparisons (whether against a fixed criterion or a norm reference), interesting “dual perspective” problems are to some extent pre-programmed, with the rater’s own attitude playing a crucial role. To take an example from child L2 acquisition: testers/raters may focus on the fact that a four-year-old girl with Russian or Turkish as L1 and German as an early L2 (first exposure at age 3), does not yet produce the range of word order patterns found in

monolingual German speakers of the same age. But how reasonable is it to even expect her to come within the range of monolinguals, given her limited exposure to German? Once examiners take into account that, despite a shorter and less intense exposure time (one year vs. four years), this child actually performs astonishingly well in comparison to her reference group, namely other children first exposed to German at the age of 3, they should be able to focus more easily on what has already been achieved, very much in the spirit of the "can do" statements of the CEFR.

Finally, let us for one moment adopt a test-taker perspective. While the areas focused on throughout the immigrant test for German are based on extensive need analyses (Perlmann-Balme et al. 2009:7) and certainly make good topics for classroom discussions outside the test situation, the test items proposed for these domains yield a depressing impression of what kind of interaction we would most likely encounter and what scary challenges we might have to cope with early on. Quite a number of items could lead us to conclude that there is considerable need for immigrants to protest, in writing, against unjustified accusations (for instance after being issued a traffic ticket; p.35), or to withdraw from a previously signed contract (p.36). We are also supposed to demonstrate our ability to request (for our children) leave of absence from school due to illness or other family catastrophe, and to apologize to teachers for having missed our language classes. There is, in short, a considerable list of either some (imagined) wrongdoing experienced or some (imagined) *faux-pas* already committed by us or imminent. Even though it is recommended that cultural clichés and stereotypes be avoided (p.27), several tasks call for comments on potential stereotypes, cf. the four instances mentioning being on time (punctuality/*Pünktlichkeit*, p. 37). Also, linguists – to step out of the test-taker role – will hardly fail to be impressed by the fact that learners are expected to comment on similarities and differences between their first language and the current target language, as well as on the usefulness of previously learned languages for the acquisition of new languages. It would seem, then, that many test demands faced by immigrant test takers are challenging enough for native speakers and even researchers, let alone for the lay person and the novice language learner.

4 CEFR futures: catching up with theory

As pointed out by McNamara, "a language test is only as good as the theory of language upon which it is based" (2000: 86). This means that tests emerging on the basis of the CEFR may be in trouble since there is no explicit linguistic theory or theoretically motivated descriptive framework backing them. This does not

just hold for grammatical form but also for functional, pragmatic aspects, and for theories of second/foreign language learning. North (2007: 27) cites 20-year-old L2 research, concluding that there is no evidence for learning orders and that linguistic theories fail to provide an uncontroversial theoretical framework for cross-linguistic comparison in the domain of grammar. Given that the search for invariant formal and functional principles and attempts at explaining cross-linguistic variation has been at the core of linguistic research for decades and within many differing theoretical approaches (e. g. Chomsky 1986; Goldberg 2006; Haegeman 1991, to mention only a few), North's conclusion appears to be unnecessarily pessimistic. While "uncontroversial" conclusions may be as unrealistic for language as for other diagnostic areas, the identification of relevant typological, formal and functional properties is possible, and could serve as a *tertium comparationis* needed for cross-linguistic comparisons.

Likewise, intensive research with respect to the acquisition types mentioned above has brought about a wealth of evidence concerning cross-linguistically comparable developmental paths and inter-individual differences. Some of the most interesting comparative data and insights were obtained in L2 research on immigrants in projects funded by the European Science Foundation in the last decades of the past century (cf. Klein and Perdue 1997; Watorek, Benazzo and Hickmann 2012). Independently of individual languages, there is strong evidence of very similar structure-building processes and correlations between the emergence of the syntactic architecture and functional categories (such as case, agreement, tense, negation; cf. Dimroth and Jordens 2009; Doughty and Long 2003; Hawkins 2001; Vainikka and Young-Scholten 2011; Watorek, Benazzo and Hickmann 2012, again to mention only a few). Regardless of theoretical persuasion, the corpora obtained within these projects and beyond, the methodological and experimental expertise, and the insight gained into learner types and developmental paths in many languages can certainly give additional impetus to attempts to place future tests based on the CEFR on a stronger theoretical and empirical footing.

5 Conclusion: chickens, eggs, and verbivores

Over the last decades, language testing, "not a topic likely to quicken the pulse or excite much immediate interest" (McNamara 2000: 3), has turned into an area of considerable relevance across all levels of the education system and across learner types and has led to intensive discussions among researchers, test designers, people in charge of language policies and allotment of resources, language teachers, therapists, and, in the case of children, parents and pediatri-

cians. While different stakeholders are not very likely to agree as to the best diagnostic instrument to use, they may be willing to concede that we should make sure that there is something to test in the first place and that it is of foremost importance to invest in the quality and quantity of the learning experience available to learners of all ages.

After all, humans are excellent language learners, and, as very pointedly formulated by Pinker, we are "verbivores, a species that lives on words" (2007: 24). As the coinage "verbivore" – a blend of "verb(al)" and "herbivore" – highlights, we get a lot of enjoyment out of the non-utilitarian functioning of our linguistic abilities. I would like to encourage those interested in optimizing language tests and especially in language teaching to make sure we do not lose sight of this resourcefulness, which comes for free – an effortless, untaught side-effect of linguistic competence. This playful function appears to be missing from current assumptions of what the immigrant "needs".

Language plays an important and ubiquitous role in identity construction, for the negotiation of common ground, for political participation and integration. While our measurements of linguistic proficiency may correlate with a number of personal traits, such as motivation, a talent for role-play, risk taking, musicality and many others, linguistic competence cannot indicate who will, in the end, be a loyal citizen, observe traffic laws and pay taxes. This means that linguists and other professionals need to draw the line and withstand increasing political pressures and demands. As pointed out by Van Avermaet (2016: 6), "These language tests often decide whether you can enter a country; stay in a country; get a permanent residency, or citizenship. They decide whether you are in or out. [...] Within this highly ideologized and politicized context, [...] language testers must reflect carefully not just on the reliability, but more than ever on the validity of their instruments."

In a way, we are faced with a chicken-and-egg problem, namely with the question of which comes first: language proficiency or integration. It is time to articulate more clearly that integration into a community of speakers who are interested in what language learners think and feel, is a most powerful incentive for language acquisition and for realizing our *verbivore* potential by exploiting the linguistic and communicative resourcefulness of learners and their interlocutors alike.

References

ALTE (Association of Language Testers in Europe). 2011. *Manual for language test development and examining*. Strasbourg: Council of Europe. http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf (accessed 17 August 2016).

ALTE. 2016. *Language tests for access, integration and citizenship: An outline for policymakers*. http://www.alte.org/attachments/files/alte_lami_language_test_policy_booklet_web_en_0 m4wh.pdf (accessed 9 August 2016).

Chomsky, Noam. 1986. *Knowledge of language: Its nature, origin, and use*. New York: Praeger.

Coste, Daniel. 2007. Contextualising uses of the Common European Framework of Reference for Languages. In F. Goullier, *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities*, 40–49. Strasbourg: Council of Europe. http://www.coe.int/t/dg4/linguistic/Forum07_webdocs_EN.asp#TopOfPage (accessed 18 August 2016).

Dimroth, Christine & Peter Jordens (eds.). 2009. *Functional categories in learner language*. Berlin: de Gruyter.

Doughty, Catherine J. & Michael H. Long (eds.). 2003. *The handbook of second language acquisition*. Malden, MA: Blackwell.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goullier, Francis. 2007. *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities*. Strasbourg: Council of Europe. http://www.coe.int/t/dg4/linguistic/Forum07_webdocs_EN.asp#TopOfPage (accessed 18 August 2016).

Grimm, Angela & Petra Schulz. 2014. Specific language impairment and early second language acquisition: The risk of over- and underdiagnosis. *Child Indicators Research* 7 (4). 821–841.

Grimm, Hannelore. 1994. Sprachentwicklungsstörung: Diagnose und Konsequenzen für die Therapie. In Hannelore Grimm & Sabine Weinert (eds.), *Intervention bei sprachgestörten Kindern: Voraussetzungen, Möglichkeiten und Grenzen*, 3–32. Stuttgart: Fischer.

Haegeman, Liliane. 1991. *Introduction to Government & Binding Theory*. Oxford: Blackwell.

Hawkins, Roger. 2001. *Second language syntax*. Oxford: Blackwell.

Hulstijn, Jan H. 2007. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91(4). 663–667.

Klein, Wolfgang & Clive Perdue. 1997. The Basic Variety (or: Couldn't natural languages be much simpler?) *Second Language Research* 13(4). 301–347.

Li Wei & Melissa G. Moyer (eds.). 2008. *The Blackwell guide to research methods in bilingualism and multilingualism*. Malden, MA: Blackwell.

McDaniel, Dana, Cecile McKee & Helen S. Cairns (eds.). 1998. *Methods for assessing children's syntax*. Cambridge, MA: MIT Press.

McNamara, Tim. 2000. *Language testing*. Oxford: Oxford University Press.

McNamara, Tim. 2011. Managing learning: Authority and language assessment. In Radhika Jaidev, Maria Luisa C. Sadorra, Wong Jock Onn, Lee Ming Cherk & Beatriz Paredes Lorente (eds.), *Global perspectives, local initiatives: Reflections and practices in ELT*,

39 – 51. Singapore: National University of Singapore, Centre for English Language Communication.
Menn, Lise & Nan B. Ratner (eds.). 2000. *Methods for studying language production*. Mahwah, NJ: Lawrence Erlbaum.
North, Brian. 2007. The CEFR Common Reference Levels: Validated reference points and local strategies. In F. Goullier, *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities*, 19 – 28. Strasbourg: Council of Europe. http://www.coe.int/t/dg4/linguistic/Forum07_webdocs_EN.asp#TopOfPage (accessed 18 August 2016).
Perlmann-Balme, Michaela, Sybille Plassmann & Beate Zeidler. 2009. *Deutsch-Test für Zuwanderer, A2–B1*. Berlin: Cornelsen.
Pinker, Steven. 2007. *The stuff of thought: Language as a window into human nature*. New York: Penguin.
Polinsky, Maria & Olga Kagan. 2007. Heritage languages: In the "wild" and in the classroom. *Language and Linguistics Compass* 1(5). 368 – 395,
Quetz, Jürgen. 2010. Der Gemeinsame europäische Referenzrahmen als Grundlage für Sprachprüfungen: Eine kritische Beschreibung des Status quo. *Deutsch als Fremdsprache* 04/2010. 195 – 202.
Strik, Tineke. 2013. Integration tests: Helping or hindering integration? Report of the Committee on Migration, Refugees, and Displaced Persons. Strasbourg: Council of Europe, Parliamentary Assembly, Doc. 13361.
Ulich, Michaela & Toni Mayr. 2003. *SISMIK: Sprachverhalten und Interesse an Sprache bei Migrantenkindern in Kindertageseinrichtungen*. Freiburg: Herder.
Vainikka, Anne & Martha Young-Scholten. 2011. *The acquisition of German: Introducing Organic Grammar*. Berlin: de Gruyter.
Van Avermaet, Piet. 2016. Foreword. In ALTE, *Language tests for access, integration and citizenship: An outline for policymakers*. http://www.alte.org/attachments/files/alte_lami_language_test_policy_booklet_web_en_0 m4wh.pdf (accessed 9 August 2016).
Vogt, Karin. 2011. *Fremdsprachliche Kompetenzprofile: Entwicklung und Abgleichung von GeR-Deskriptoren für Fremdsprachenlernen mit einer beruflichen Anwendungsorientierung*. Tübingen: Narr.
Watorek, Marzena, Sandra Benazzo & Maya Hickmann (eds.). 2012. *Comparative perspectives on language acquisition: A tribute to Clive Perdue*. Bristol: Multilingual Matters.
Wisniewski, Katrin. 2010. Bewertervariabilität im Umgang mit GER-Skalen: Ein- und Aussichten aus einem Sprachtestprojekt. *Deutsch als Fremdsprache* 03/2010. 143 – 149.