EURAC research

The variable interpretation of CEFR scales

Which components of CEFR scales do raters of different language backgrounds report to focus on when rating writing tests?

Preliminary remarks: Synergies and tensions regarding language matters in South Tyrol

- Situation of language contact: physical closeness ≠ readiness for dialogue (Baur 2000); friendships mostly inside the own language group, private life is monolingual … (cfr. Language Barometer 2006) = mostly monolingually oriented separate social groups

- L2 competence not satisfactory (Putzer/Deflorian 1997, Egger 2001, CENSIS 1997, ASTAT 2007, Vettori 2005, Language Barometer 2006) → mostly self-evaluations

1. The KOLIPSI project – background information

- *European Academy of Bolzano* (EURAC), *Department of Cognitive and Educational Sciences* in Trento, German and Italian Education Authorities
- Aim:
  - (1) extensive analysis of the L2 writing (and oral) competences (productive and interactive communicative activities);
  - (2) sociolinguistic and psychosocial analysis of extra-linguistic factors of influence on competence level
- **Test takers**: appr. **1.200** pupils(17-18 years old)
- Instruments: questionnaires, interviews, focus groups, diff. **language tests**

→ *KOLIPSI writing test*

2. KOLIPSI writing test

- productive and interactive communicative activities
- L2 = Italian + German
- targel levels: B1/B2
- two tasks: email / letter
- relation to CEFR tried to be established

Quality management (raters):

- rater training
- regular follow-up meetings
- double blind rating of all tests
- external raters (5%)
- variety of quantitative procedures to guarantee reliability, control rater effects … (IRT procedures used)

3. The study

Research questions:

- What elements of CEFR scales do trained raters report to focus on when rating writing tests?
- Are there any differences between the behaviour of Italian and German speaking raters?

3. The study

Prior research…

- Studies concentrating on mental processes
- Studies concentrating on "rater styles" / features in focus

→ *KOLIPSI: focus on components/key features of CEFR scales (grammatical accuracy, vocabulary, coherence, sociolinguistic appropriateness)*

3. The study

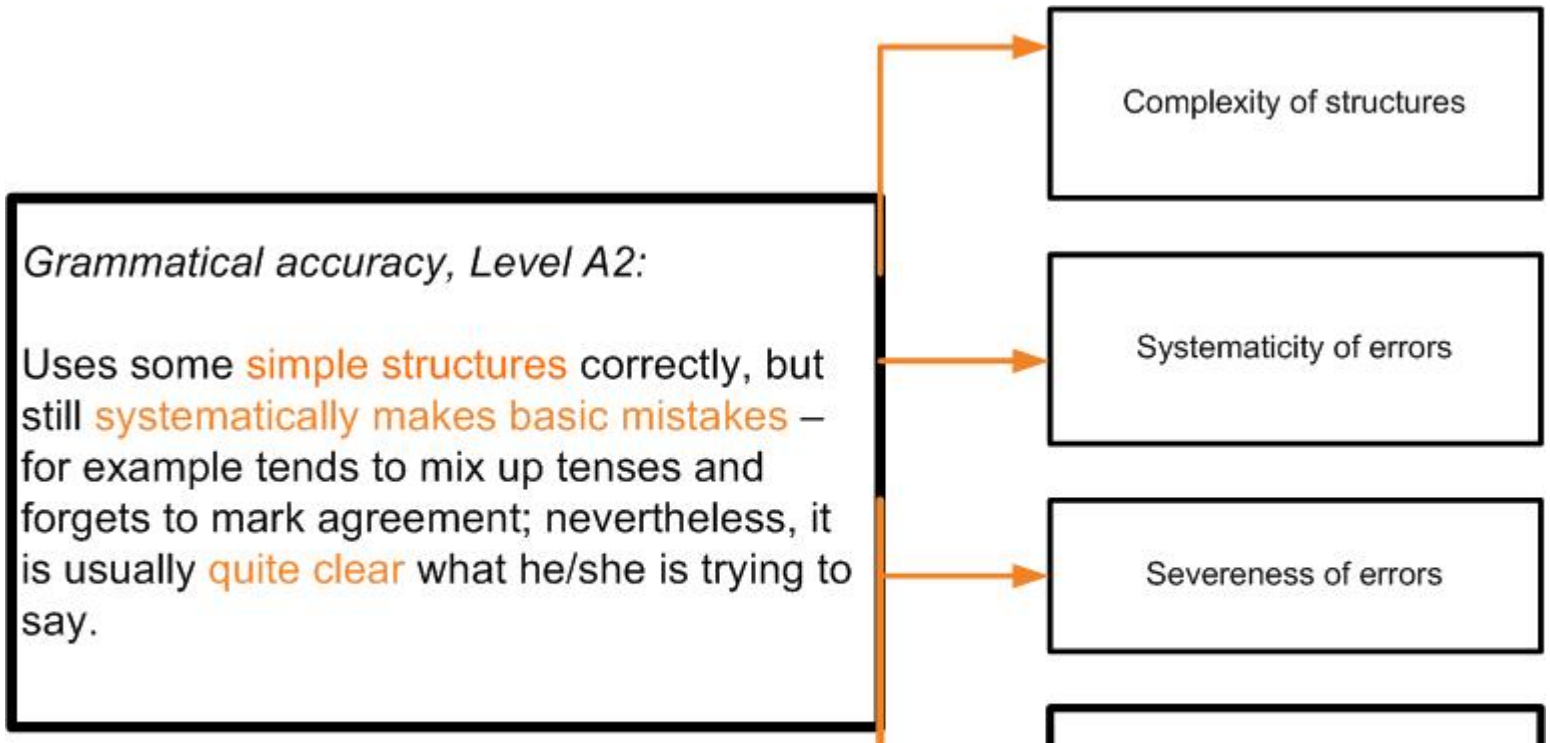
The raters

- 6 Italian, 3 German raters
- mostly language teachers
- used rating scale based on slightly modified CEFR scales (levels A1-C1, no intermediate levels (A2+, B2+))

Method

- Rater **questionnaire**, focussing on …
  - difficulty handling rating criteria
  - focus features in decision making
    → **extracted and isolated aspects of CEFR scales (+ "distractors")**
  - problems with formulations in CEFR scales
  - other, project- and CEFR-related issues

3. The study: construction of the questionnaire



3. The study

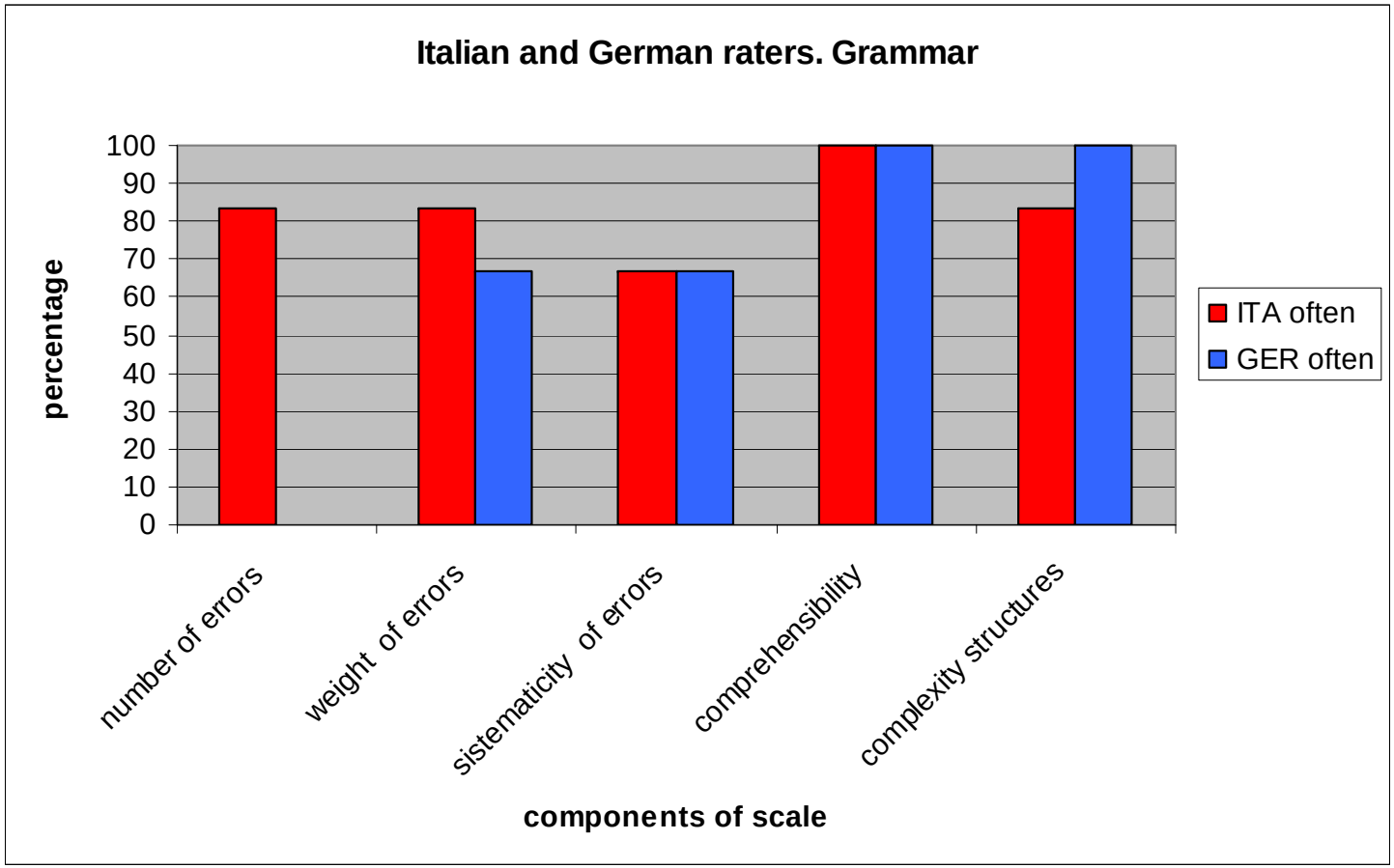
Data analysis

- qualitative analysis of answers to open-ended questions
- descriptive, very simple quantitative analysis

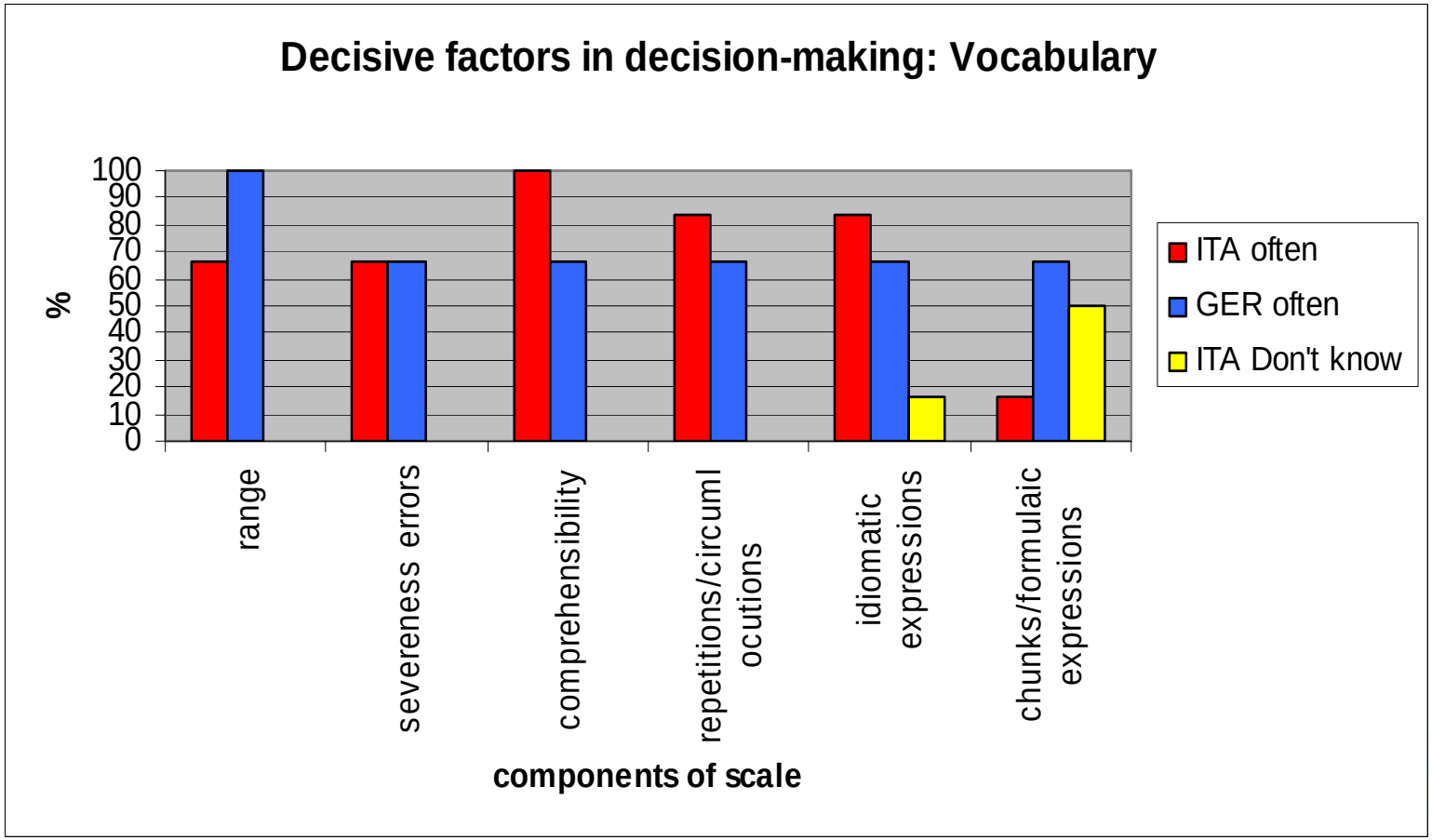
Constraints…

- No generalisations possible: small database!
- Exploratory character; generation of hypotheses

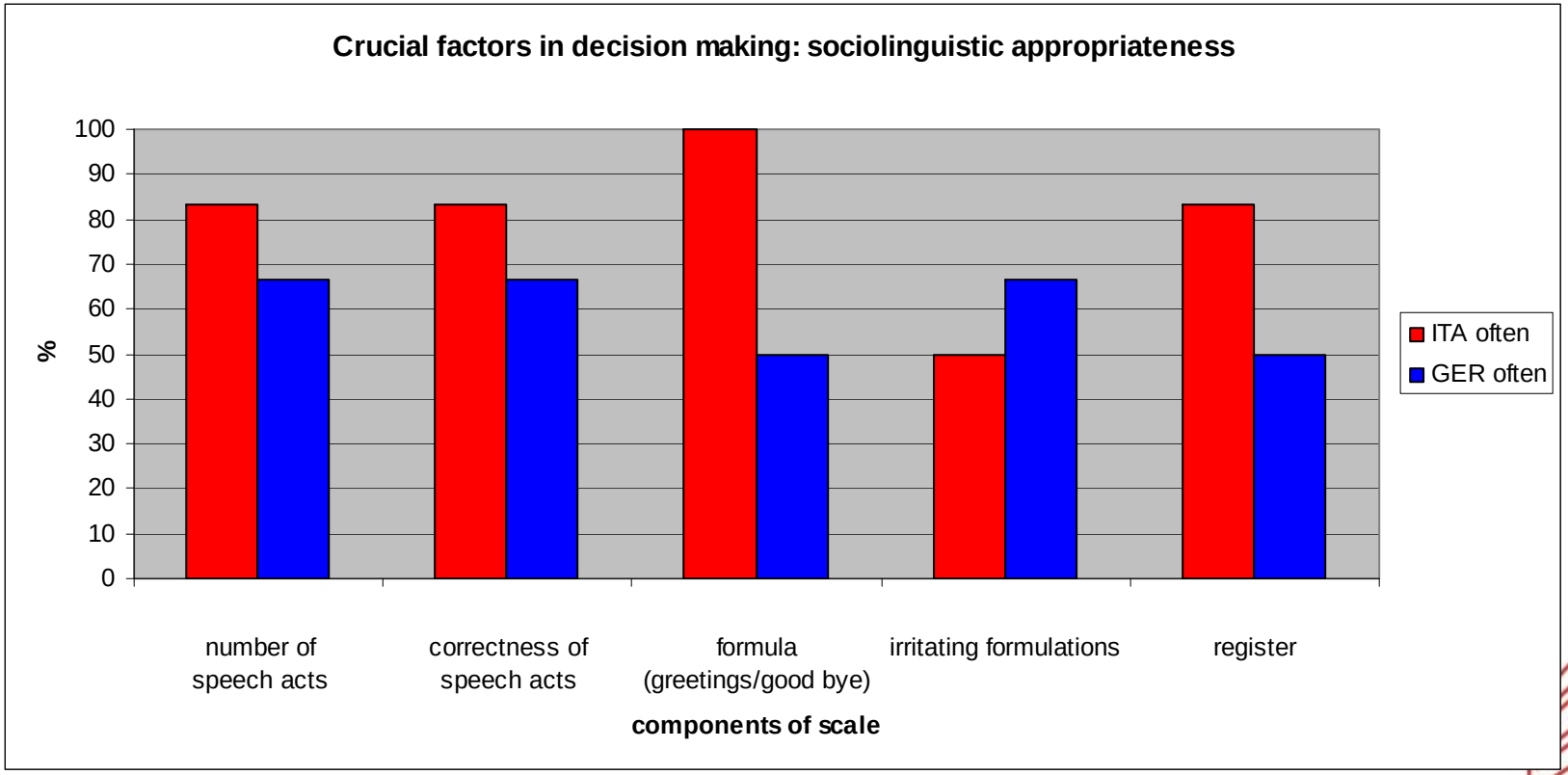
4. Results: Grammatical accuracy



4. Results: Vocabulary



4. Results: Sociolinguistic appropriateness



Summary of findings

1. Raters attribute strongly varying weight to concepts of scales - inside and across rater language groups.
2. Raters use concepts that are NOT in the scales.
3. Raters overgeneralise single descriptors to whole scales, especially when they are particularly “handy“ and concrete.
4. Even after training, raters feel unsure how to use some formulations in scales.

Conclusion

1. Possible reasons for variation…
   - lack of training (?)
   - lack of sensitivity of raters towards scales
   - language or cultural background/testing cultures
   - individual concepts of language competence

2. Next steps…
   - EMPIRICAL CHECK → What do raters REALLY focus on?
     → *learner corpus of KOLIPSI writing tests*

Thank you for your attention!

References

ASTAT 2007 = Autonomous Province of Bolzano/South Tyrol-Provincial Institute for Statistics (2007) *South Tyrol in Figures.* Bolzano – Online document: http://www.provinz.bz.it/downloads/Siz_2007-eng.pdf.Baur, S. (2000) *Die Tücken der Nähe. Kommunikation und Kooperation in Mehrheits-/Minderheitssituationen.* Bozen: alpha&beta.

Brown, Annie/Iwashita, Noriko/McNamara, Tim (2005): *An Examination of Rater Orientations and Test Taker Performance on English for Academic Purposes Speaking Tasks*. Princeton, NJ: Educational Testing Service.

Censis/Autonome Provinz Bozen (1997): Identität und Mobilität der drei Sprachgruppen in Südtirol. Abschließender Bericht. Rom.

Cumming, A./Kantor, R./Powers, D.E. (2001): *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision-making and development of a preliminary analytic framework.* (TOEFL Monograph No. MS-22). Princeton, NJ:ETS. FL ok 13.3.

Cumming, A./Kantor, R./Powers, D.E. (2002): Decision-making while rating ESL/EFL writing tasks: A descriptive framework. In: *The Modern Language Journal* 86, 67-96.

Eckes, Thomas (2008): Rater types in writing performance assessments: A classification approach to rater variability. In: Language Testing 25 (2), 155-185.

Hamp-Lyons, L. (ed.) (1991): *Assessing Second Language Writing in Academic Contexts.* Norwood, N.J.: Ablex Publishing Corporation.

Language Barometer= ASTAT 2006, Autonomous Province of Bolzano/South Tyrol-Provincial Institute for Statistics (2006) *Südtiroler Sprachbarometer 2004. Sprachgebrauch und Sprachidentität in Südtirol*. Bozen: Autonome Provinz Bozen/Südtirol.

Lumley, T. (2005): *Assessing Second Language Writing: The Rater's Perspective*. Frankfurt a.M. u.a.: Lang.

Milanovic, M./Saville, N./Shuhong, Shen (1995): A study of the decision-making behavior of composition markers. In: Milanovic, M./Saville, N. (eds.): *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium. Studies in Language Testing 3*. Cambridge: CUP, 92-114.

Putzer/Deflorian, F. (1997): *Considerazioni riassuntive sui risultati delle prove nella comprensione scritta e orale, nella produzione scritta e nella produzione orale della lingua seconda degli alunni delle ultime classi della scuola elementare, media e superiore in lingua tedesca*. Bolzano: Provincia Autonoma di Bolzano, Ufficio Bilinguismo (unpublished)

Vaughan, C. (1991): Holistic assessment: What goes on in the rater's mind? In: Hamp-Lyons.

Vettori, Chiara (2005) *La competenza del tedesco degli studenti italofoni di scuola media inferiore e superiore di Bolzano e Trento: confronto e valutazione. Tesi per il conseguimento del titolo di dottore di ricerca.* Università di Modena e Reggio Emilia. Weigle, S.C. (1994): Effects of training on raters of ESL compositions. In: *Language Testing* 11(2), 85-106.

Pollitt, A./Murray, N.L. (1996): What raters really pay attention to. In: Milanovic, M./Saville, N. (eds.): P*erformance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium.* Cambridge: CUP, 74-91.