Towards Reasoned Modality Selection in an Embodied Conversation Agent

Carla Ten-Ventura[1], Roberto Carlini[1],
Stamatia Dasiopoulou[1], Gerard Llorach Tó[1], and Leo Wanner[1,2]✉

[1]Universitat Pompeu Fabra, [2]ICREA
Barcelona, Spain
email: first_name.last_name@upf.edu

Abstract. We present work in progress on (verbal, facial, and gestural) modality selection in an embodied multilingual and multicultural conversation agent. In contrast to most of the recent proposals, which consider non-verbal behavior as being superimposed on and/or derived from the verbal modality, we argue for a holistic model that assigns modalities to individual content elements in accordance with semantic and contextual constraints as well as with cultural and personal characteristics of the addressee. Our model is thus in line with the SAIBA framework, although methodological differences become apparent at a more fine-grained level of realization.

1 Introduction

In order to appear natural and thus be accepted by human interlocutors, embodied conversation agents are expected to appropriately use language, facial expressions and gestures. A considerable number of works addresses the two aspects of the problem: (i) when to select what modality, and (ii) how to synchronize the different modalities such that the overall (verbal and non-verbal) behavior of the agent appears coherent and natural. In the recent past, the problem has often been restricted to planning of the non-verbal behavior of an agent [7, 19]. In this case, the verbal mode is assumed to be already given, either as speech (i.e., acoustic stream) [1, 12] or in terms of written statements [20]. To plan the facial expressions and gestures, the speech respectively written statements are then analyzed and, depending on the identified linguistic and/or content features [18, 4, 5], specific facial expressions and gestures are assigned to acoustic / linguistic (word sequence) segments. While this strategy seems appropriate when an off-the-shelf verbal communication generator is used or when the agent is supposed to follow a predefined already spelled-out script, it is counter-intuitive from a holistic perspective on dynamic communication: facial expressions and gestures are not simply an add-on to language. Rather, as argued in theoretical studies [10, 13, 14, 9] and as already assumed in the early days of the research on conversation agents [3], all modalities play together in order to produce a natural communication move of the agent. For instance, in an affirmative act, the agent may nod, smile and say *Yes, that's correct!* or simply nod; to indicate

a location, it may say *Over there* and/or produce a deictic gesture; to express an intense rejection, it may say *I don't like it.* and signal via a facial expression the intensity, or choose the verbal mode to communicate the intensity as well (*I don't like it at all!*); and so on. Such a holistic view on the planning of a move is required, for instance, in the context of a flexible embodied multilingual and multicultural conversation (i.e., dialogue) agent as targeted in the KRISTINA Project. This agent (henceforth referred to as "KRISTINA") is expected to be able to flexibly act in different contexts as a basic care assistant, health care adviser or social companion of humans; see [23] for an overview.

In multimodal dialogue and virtual agent research, several proposals have been made towards a holistic *fission* model. Cf., e.g., [6, 22] for proposals on the dialogue side, which tend to assign a specific modality or a combination thereof to moves or to move elements in predefined dialogue scripts. The problem with these proposals is that when broader conversation topics are to be covered and the agent needs to count with spontaneous interventions of the human (as is our case), predefined dialogue scripts are not adequate. In the context of virtual agent research, the most influential proposal has been the SAIBA-framework[1]. SAIBA foresees three stages of behavior realization (see, e.g., [21]): Intent Planning, Behavior Planning, and Behavior Realization. Modality selection is foreseen to take place in the Behavior Planning (BP) module. However, BP has to span between abstract *Functional Markup Language* communicative intention representations as output by the Intent Planning (see [2] for examples) to a very detailed synchronization alignment between specific modality realizations. We believe that it is necessary to separate modality assignment from synchronization of the specific modality realizations since both tasks are situated at very different levels of abstraction and require different types of information.

In what follows, we present work in progress on a holistic versatile modality selection model that is embedded into the multimodal dialogue architecture of the KRISTINA agent. Section 2 situates modality selection in this architecture. In Section 3 then our approach to modality selection and realization is discussed. Section 4, finally, draws some conclusions and discusses our ongoing and future work on modality selection.

2 Modality Selection in KRISTINA

Figure 1 displays the part of the KRISTINA architecture into which our modality selection model (marked in the figure by a box) is embedded. KRISTINA is a knowledge-based agent. The semantic structures produced by the multimodal communication analysis modules (not shown in Figure 1) are projected onto genuine ontological (OWL) structures, fused and integrated by the knowledge integration (KI) service into the knowledge base (KB). Furthermore, a dedicated search engine feeds into the KI service background multimodal information from the web and relevant curated information sources in order to ensure that the

[1] `http://www.mindmakers.org/projects/saiba/wiki`

agent is knowledgeable about the topics raised by the human counterpart and to facilitate the realization of flexible reasoning-based dialogue strategies.

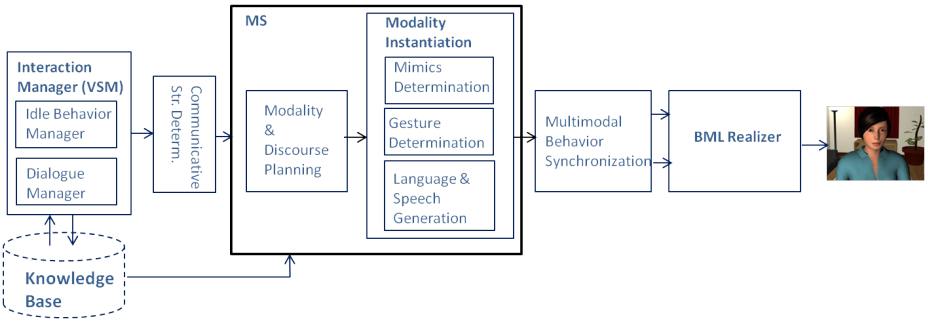

Fig. 1. Situating modality selection in the KRISTINA architecture

Modality Selection (MS) receives input from two sources: the modules controlled by the Interaction Manager and the addressee (or human conversation counterpart) profile, which is stored in the KB. The Interaction Manager is embedded in the *Visual Scene Maker* (VSM) framework [8]. While the original purpose of VSM has been to support the definition of the interactive behavior of virtual characters, we use it, on the one hand, as a communication shell between the dialogue management module and the modules it interacts with, and, on the other hand, for modeling the idle behavior of the agent. The dialogue manager (DM) chooses the best system reaction (in terms of ontological structures), in accordance with the analyzed user move, the user's emotion and culture and the recent dialogue history. For this purpose, it solicits first from the KI module possible reactions that are reasoned over the KB. In other words, in contrast to most of the state-of-the-art DM models, the determination of the turn of the system is distributed between a high level control DM and a reasoning KB module; see [15] for details. On its way to the MS module, the DM output is enriched by *communicative* labels. The communicative labels mark which parts of the content are considered to be the core of the trasmitted message, which are to be highlighted, which are to be backgrounded, etc. Their distribution depends on the communicative intention of the agent. So far, we mark only the core of the message and the "topic" on which this core elaborates. Note that in linguistic terms, the core of a message is referred to as *rheme* (the content that the speaker aims to transmit, i.e., the essential part) and the topic of the core as *theme* (to what the transmitted content refers). For instance, in the statement *You look worried today*, the theme is *You* (the statement is about 'you') and rheme is *look worried today*. Theme and rheme are reflected by prosody and body gestures of a speaker. Several works on behavior modeling thus analyze the given verbal part of an agent move to detect theme and rheme in order to introduce, e.g.,

gestures or pitch accents to mark the rheme; see, e.g., [18, 20, 4]. Consider, for illustration, an example of the APML codification of rheme, in this case, in the early Greta agent in Figure 2: it is assigned to the string *Good morning Mr. Smith*, which is further enriched by prosodic markers.

```
<turnallocation type="take"> <performative type="greet">
<rheme>Good<emphasis x-pitchaccent="Hstar">morning
</emphasis> Mr Smith.<boundary type="LL"/></rheme>
</performative></turnallocation>
```

Fig. 2. Codification of a sentence in APML (example taken from [4])

This retrospective analysis is obsolete in our design. Furthermore, the assignment of communicative labels to content elements (instead of linguistic constructions) has the advantage that they can be used and realized by each individual modality independently, by means that are available to this modality. For instance, in the case of the verbal modality, the theme/rheme tags are used to shape and later linearize the syntactic structure and to derive prosodic markers. In the case of gestures, they are used, e.g., to determine beat gestures.

Figure 3 shows a sample input structure provided to MS, i.e., the DM output structure after it has been enriched by the thematicity tags.[2]

```
:sa1 a la:SystemAction ;
   dialogue:contains :da1 .
:da1 a dialogue:Declare ;
   dialogue:arousal 0.5 ;
   dialogue:valence 1.0 .
   dialogue:semContent :sit1, :sit2.
:sit1 a la:Situation ;
   la:contains :stmt1, :stmt2 .
:sit2 a la:Situation ;
   la:contains :stmt2, :stmt3 .
:stmt1 la:context (:like1 a onto:Like) ;
   la:agent (:Alp a onto:CareRecipient).
:Alp thematicity:theme true.
:stmt2 la:context :like1 ;
   la:theme (:ins1 a onto:Baklawa) .
:ins1 thematicity:rheme true.
:stmt3 la:context :like1
   la:manner (:ins2 a onto:Always);
:ins2 thematicity:rheme true.
:like1 thematicity:rheme true.
```

Fig. 3. Input structure to Modality Selection

The structure contains the following types of information: 1. name of the dialogue act (`Declare`), 2. the content that is to be communicated by the agent (under 'dialogue:semContent'), 3. valence (of the agent), 4. arousal (of the agent), and 5. thematicity (theme/rheme) labels. It encodes the facts that a care recip-

[2] Note that we use the Turtle notation (`https://www.w3.org/TR/turtle/`), such that, e.g., ":da1 a dialogue:Declare" means that ':da1' is an instance of the dialogue act class 'Declare').

ient Alp likes Baklawa, and that it always used to like it. Alp is thus the theme and the other content elements constitute the rheme.

From the addressee (or user) profile, MS uses a series of features: culture to which the addressee belongs (Central European, South European, Northern, ...), age, gender, personality (extroverted or introverted), proximity to the agent (close, familiar, or distant), etc. This allows us to adapt the communication of the agent to its conversation counterpart, e.g., in terms of the quantity, distribution and type of gestures and mimics.

Modality Selection (MS) is performed in KRISTINA in two stages. In the first stage, the modalities are first assigned to the content elements in the received input structure (note that a structure can consist of one single element, and that to a given element more than one modality can be assigned) and then related in terms of a discourse structure.[3] In the second stage, the modalities are instantiated, i.e., for each modality it is determined how it will be realized (smile, head turn, specific verbalization, etc.). The first stage is processed by the *Modality & Discourse Planning* module; the second stage by the *Modality Instantiation* module.

The output of the Modality Instantiation module is fed into the *Behavior Synchronization* module, which is the lean version of the *Behavior Planner* in the sense of the SAIBA-framework (it focuses only on the synchronization of the modalities) and, which, in its turn, passes its output to the BML realizer (again, in the sense of SAIBA). Let us focus now, however, on the two stages of modality selection.

3 Getting the Multimodal Message Across

Prior to the choice of a specific realization of a modality to express some content, as, for instance, *Hello!* (rather than *Good evening, Sir!*) for the verbal modality of greeting, or head shaking for the gesture modality of negation (either to emphasize the verbal *No* or as a sole act), we must first choose the appropriate modalit(y/ies) for each content element provided by the DM. The nature of both types of choices is rather different, as far as cognitive and communicative criteria are concerned. Furthermore, it is desirable from the theoretical[4] and methodological viewpoints to treat language, gesture and mimics generation analogously. Thus, in language generation, a content structure is projected (preferrably in a cascade of transitions) onto one of its possible linguistic realizations in accordance with linguistic means that are not known by the modality selection. In the same way, a content structure should be projected, e.g., onto a concrete facial expression respectively gesture by separate modules, without that modality selection has to know how rich the facial expression/gesture repertoire of the facial

[3] While the dialogue manager plans the structure of the discourse between dialogue acts, it does not intervene in the definition of the internal discourse structure of a dialogue act/move because this would require a "look-ahead" for the acquisition of information on the realization of the move.

[4] Cf. the references to the theoretical studies in the Introduction

expression/gesture realizer (in our case, the virtual character) is. Therefore, we deal with modality selection and modality determination separately in different submodules. On the other hand, when we decide, for instance, that a greeting is to be expressed by a verbal statement and a gesture, we need to already relate both expressions to each other (whether one is to be considered as an emphasis of the other, as an "elaboration", etc.). In other words, we need to establish the internal discourse structure of the move of the agent. As a consequence, we deal with the modality assignment and internal discourse structure determination in one module.

3.1 Modality and discourse planning

The output of the *Modality & Discourse Planning* module looks as shown in Figure 4

```
:sa1 a la:SystemAction ;
   dialogue:contains :da1, fe1 ; .
   ms:verbal :da1 .
   ms:facialExpression :fe1 .
:da1 a dialogue:Declare ;
   dialogue:arousal 0.5 ;
   dialogue:valence 1.0 .
   dialogue:semContent :sit1, :sit2.
:sit1 a la:Situation ;
   la:contains :stmt1, :stmt2 .
:sit2 a la:Situation ;
   la:contains :stmt2, :stmt3 ;
   rst:restatement :sit1 .
:stmt1 la:context (:like1 a onto:Like) ;
   la:agent (:Alp a onto:CareRecipient).
:Alp thematicity:theme true.
:stmt2 la:context :like1 ;
   la:rheme (:ins1 a onto:Baklawa) .
:ins1 thematicity:rheme true.
:stmt3 la:context :like1
   la:manner (:ins2 a onto:Always);
:ins2 thematicity:rheme true.
:like1 thematicity:rheme true.
:fe1 a ms:JoyfulExpression ;
   ms:hasIntensity "high" ;
   rst:parallel :da1 .
```

Fig. 4. Output structure of Modality Selection

As we can observe, the modalities have already been assigned to the content elements (cf. 'ms:verbal :da1' and 'ms:facialExpression :fe1'). In what follows, we outline how this is achieved.

Modality planning. The assignment of the modalities to the individual content chunks is currently rule-based. Consider, for instance, a fragment of a rule, formulated for transparencey in pseudo-code XML format in Figure 5. This rule assigns to the whole dialogue act an intense joyful facial expression (which will be mapped during the determination of the facial expression onto a broad smile; see below) if, for instance, KRISTINA tells a care giver that the elderly Turkish person in question, who is from the region of Ankara, likes Baklawa. Note that in order to deduce the required information, the agent needs to reason.

```
<conditions>
    <da>Declare<id>'id1'</id> </da>
    <topic>eating_habits</topic>
    <CareRecipient><age>elderly></CareRecipient>
    <theme> type(theme) == 'food' ∧ type(food) == 'traditional' ∧
        region(food) == origin(CareRecipient)</theme>
</conditions>
<modality> <id>'id1'</id>
    <fe> JoyfulExpression <intensity>high</intensity></fe>
    <valence> valence(id1)</valence><arousal>arousal(id1)</arousal></fe>
</modality>
```

Fig. 5. Fragment of a mode selection rule

Intra-move multimodal discourse structure planning. Given that in KRISTINA there is no "ground" modality (as, e.g., language in many of the previous works) to which then the other modalities are assigned (and thus synchronized), but, rather, all three modalities are used as equal and assigned to content elements in the same dialogue act quasi independently from each other, they need to be related in order to form a coherent discourse. This is especially of relevance if a dialogue act contains several statements (see also Footnote 1 on the competence of the dialogue manager). Then, apart from the discourse alignment between the modalities, a discourse structure between the verbal elements must be defined. For this purpose, we explore a discourse structuring technique that originates from text generation [17]. The technique is based on the *Rhetorical Structure Theory* (RST) [11]. Apart from the conventional set of RST relations (such as ELABORATION, CAUSE, JUSTIFICATION, etc.), the relation SIMULTANEITY is to be used. The relations hold between *elementary discourse units* (EDU) (usually, individual facts) to which one or several modalities are assigned. In the output structure above, the discourse relation tag is introduced as 'rst:parallel'.

3.2 Modality Instantiation

As described above, modality selection determines the modality of each content element or EDU, but it does not determine the specific implementation of the modality, i.e., it does not instantiate it. For instance, in the rule example above, it is determined that the facial expression has to be a intense and joyful, but not that it is a broad smile. Each modality is instantiated separately and then passed to the *Multimodal Behavior Synchronization* module, where the instantiations in different modalities are synchronized in terms of the *Behavior Markup Language* (BML) [21], drawing upon the temporal conditions imposed by the relations of the RST discourse structure. The output is a BML description that is passed to the BML Realizer, where the instantiated gestures and facial expressions are generated, in synchrony with the language uttered by the agent.

The realization of the verbal modality is carried out by a full-fledged multilingual text generator [16]. The facial expressions and gestures that can be

handled by the character are specified in the so-called *mimicon* respectively *gesticon*, where to each facial expression / gesture its high level description features as provided by the MS are assigned. Consider, for illustration, a sample entry in the *mimicon* Figure 6.

```
<description>
    <fe>JoyfulExpression<intensity>high</intensity></fe>
    <valence>1.0</valence><arousal>0.5</arousal></description>
<mimics> broad_smile</mimics>
```

Fig. 6. Sample entry of the *mimicon*

4 Conclusions and Future Work

We have presented work in progress on dynamic modality selection in embodied conversational agents. Being dynamic, i.e., guided by contextual, content and addressee profile features, it is different from most of the approaches to modality handling in multimodal dialogue systems, which tend to assign modalities *a priori* to predefined dialogue scripts. At the first glance, it is similar to the design of the *Plan Enricher* in the MagiCster project [4] in that it receives its input structure from the dialogue manager and draws upon a knowledge base. However, unlike the Plan Enricher, which provides an APML structure in which, e.g., the verbal statements are already predefined, and the mimicry fully spelled out and synchronized, we delegate language generation to a dedicated language generator and the mimicry and gestures realization to the BML Realizer. We also separate modality selection from intra-move multimodal discourse structure planning and modality instantiation. This has the advantage that the model is more generic. Our model can be considered as a proposal for an alternative realization of the Behavior Planner in the SAIBA-framework. Instead of dealing with the problem of the projection of very abstract communicative intention representations onto specific behavior realizations, we propose to divide the problem into a series of subproblems, each of which is dealt with in a separate submodule: (i) modality selection, (ii) discourse planning, (iii) modality instantiation, and (iv) modality synchronization.

Our illustrations draw upon the current rule-based prototypical implementation of the module. This implementation takes so far only a limited number of contextualized conditions into account and makes only limited use of the reasoning and inference potential of KRISTINA's reasoning engine. Also, it ignores the fact that several rules that target the selection of the same modality may overlap in their conditions and thus lead to a conflict during modality realization. This is excluded within the current simplified model, but cannot be ruled out in a more complex model. In the future, we plan to complete the development of the

proposed model, including a mechanism for rule conflict resolution, and also to learn modality selection using supervised learning techniques. For this purpose, we are in the process of annotating a corpus of multimodal spoken conversation recordings with modality and valence/arousal information. Furthermore, a quantitative and qualitative evaluation is about to be carried out in order to assess the performance of our modality selection strategy compared to the state of the art.

Acknowledgments

The presented work is funded by the European Commission as part of the H2020 Programme, under the contract number 645012–RIA. Many thanks to the three reviewers for their very helpful comments and suggestions.

References

1. Albrecht, I., Haber, J., Seidel, H.P., Earnshaw, R.: Automatic generation of non-verbal facial expressions from speech. In: Proceedings of the International Computer Graphics Conference. pp. 283–293 (2002)
2. Cafaro, A., Vilhjalmsson, H., Bickmore, T., Heylen, D., Pelachaud, C.: Representing Communicative Functions in SAIBA with a Unified Function Markup Language. In: Intelligent Virtual Agents. pp. 81–94. Springer Verlag, Heidelberg (26 Aug 2014)
3. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjámsson, H., Yan, H.: Embodiment in conversational interfaces: Rea. In: Proceedings of CHI '99. pp. 520–527. ACM (1999)
4. De Carolis, B., Pelachaud, C., Poggi, I., Steedman., M.: APML, a mark-up language for believable behavior generation. In: Prendinger, H., Ishizuka, M. (eds.) Lifelike Characters. Tools, Affective Functions and Applications. Springer Verlag (2004)
5. Endrass, B., Rehm, M., André, E.: Planning Small Talk behavior with cultural influences for multiagent systems. Computer Speech and Language 25, pages = (2014)
6. Foster, M.: Interleaved preparation and output in the comic fission module. In: Proceedings of the ACL Workshop on Software. Ann Arbor (2005)
7. Freigang, F., Kopp, S.: This is whats important – using speech and gesture to create focus in multimodal utterance. In: Lecture Notes in Computer Science, pp. 96–109 (2016)
8. Gebhard, P., Mehlmann, G.U., Kipp, M.: Visual SceneMaker: A Tool for Authoring Interactive Virtual Characters. Journal of Multimodal User Interfaces: Interacting with Embodied Conversational Agents, Springer-Verlag 6(1-2), 3–11 (2012)
9. Kendon, A.: Gesture. Visible action as utterance. Cambridge University Press, Cambridge (2004)
10. Lock, A. (ed.): Action, gesture, and symbol: The emergence of language. Academic Press, London & New York (1978)
11. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text - Interdisciplinary Journal for the Study of Discourse 8(3), 243–281 (November 2009)

12. Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., Shapirok, A.: Virtual character performance from speech. In: SCA '13 Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. pp. 25–35 (2013)
13. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press, Chicago (1992)
14. McNeill, D. (ed.): Language and gesture. Cambridge University Press, Cambridge (2000)
15. Meditskos, G., Dasiopoulou, S., Pragst, L., Ultes, S., Vrochidis, S., Kompatsiaris, I., Wanner, L.: Towards an Ontology-Driven Adaptive Dialogue Framework. In: Proceedings of the 1st International Workshop on Multimedia Analysis and Retrieval for Multimodal Interaction (MARMI). pp. 15–20. ACM, New York (2016)
16. Mille, S., Burga, A., Carlini, R., Wanner, L.: FORGe at SemEval-2017 Task 9: Deep sentence generation based on a sequence of graph transducers. In: Proceedings of SemEval '17. Association for Computational Linguistics, Vancouver (2017)
17. Moore, J., Paris, C.: Planning Text for Advisory Dialogues. Capturing Intentional and Rhetorical Information. Computational Linguistics 19(4), 1–46 (1993)
18. Pelachaud, C., Badler, N.I., Steedman, M.: Generating facial expressions for speech. Cognitive Science 20, 1–46 (1996)
19. Quintas, J., Menezes, P., Dias, J.: Auto-Adaptive interactive systems for active and assisted living applications. In: IFIP Advances in Information and Communication Technology, pp. 161–168 (2016)
20. de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., De Carolis, N.: From Greta's Mind to her Face: Modeling the Dynamics of Affective States in a Conversational Embodied Agent. International Journal of Human-Computer Studies 59(1–2), 81–118 (2003)
21. Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A., Pelachaud, C., Ruttkay, Z., Thórisson, K., van Welbergen, H., van der Werf, R.: The behavior markup language: Recent developments and challenges. In: Intelligent Virtual Agents. pp. 99–111. Springer Verlag, Heidelberg (17 Sep 2007)
22. Walker, M., Whittaker, S., Stent, A., Maloor, P., Moore, J., Johnston, M., Vasireddy, G.: Generation and evaluation of user tailored responses in multimodal dialogue. Cognitive Science 28(5), 811–840 (2004)
23. Wanner, L., André, E., Blat, J., Dasiopoulou, S., Farrùs, M., Fraga, T., Kamateri, E., Lingenfelser, F., Llorach, G., Martínez, O., Meditskos, G., Mille, S., Minker, W., Pragst, L., Schiller, D., Stam, A., Stellingwerff, L., Sukno, F., Vieru, B., Vrochidis, S.: KRISTINA: A Knowledge-Based Virtual Conversation Agent. In: Demazeau, Y., Davidsson, P., Vale, Z., Bajo, J. (eds.) Advances in Cyber-Physical Multi-Agent Systems. The PAAMS Collection – 15th International Conference, PAAMS 2017. Springer, Heidelberg (2017)