Co-Occurrences of Context Dimensions of Spreadsheets

Andrea Kohlhase[1] and Ana Guseva[2]

[1] University of Applied Sciences Neu-Ulm
D-89231 Neu-Ulm, Germany
Andrea.Kohlhase@hs-neu-ulm.de
[2] Macedonia

Abstract. We are interested in the spreadsheet as a form of mathematical user interface – human-spreadsheet interaction. Concretely, we aim for a better understanding of spreadsheet comprehension. We look into the "context space" of spreadsheets, i.e., the space induced by information dimensions along which users try to grasp the presented content. Our recent research has shown that, on the one hand, there is a clear difference in the context spaces of spreadsheet readers and authors. On the other hand, spreadsheet complexity does not distinguish the users' context space.
In this paper we look even deeper into the elicited data to find out whether and if so, how context dimensions depend on each other. If there are significant differences between such co-occurrences with respect to user roles or complexity, then we can make use of the results to enhance user-assistance systems for spreadsheets.

Keywords: Spreadsheets; repertory grid; human-spreadsheet interaction; information objects

1 Introduction

Spreadsheets have become very popular tools for analyzing and visualizing data from business and science. It has been estimated that each year tens of millions professionals and managers create hundreds of millions of spreadsheet programs [Pan00]. This intensity yields not only more and more shared, complex spreadsheet programs, but also wide-impact errors on the data level (up to 90% [Pan00], see also [PLB08]) and on the comprehension level (e.g. [PBL08]). The field of human-spreadsheet interaction deals with the processes and objects of interaction of humans with spreadsheet programs. It is one approach to tackle the failure of spreadsheets.

The bulk of research addressing this proposes more rigorous and guided practices for spreadsheet programming or addresses spreadsheet auditing, i.e., supporting practitioners with debugging existing spreadsheets for errors – see e.g. the bibliography of the European Spreadsheet Risks Interest Group (EusSpRiG) conference series at `http://www.eusprig.org/`. All of these only address the spreadsheet user in her role as the "spreadsheet author", who either is assisted at the time of initial spreadsheet creation or maintenance. Notable exceptions are [HG93; Koh10; HG94; Wol+11] who base their research on NARDI and MILLER's work on spreadsheets as multi-user applications [NM90a]. In particular, these authors consider spreadsheets as communication

and collaboration tools to exchange or combine domain knowledge and coding expertise.

But according to [Bak+08; CS10; CMW07; SSM05; HG94; NM90b] the use of spreadsheets includes the following use cases as well:

- making use of existing templates by simply putting in new data,
- reviewing data developments on different abstraction levels e.g. supervisors or members of a board,
- assessing data to base further decisions upon,
- re-understanding spreadsheet program after a period of non-use, or
- searching for reusable parts of a spreadsheet program, therefore browsing available ones.

Hence, spreadsheet use cannot be reduced to spreadsheet authoring. Our approach is to explicitly differentiate spreadsheet users into authors and readers. We have started to study the distinction between spreadsheet readers and authors in our recent research. For instance, our results show that *a*) spreadsheets only convey "information" and not "knowledge" [Koh13], *b*) the context of information by readers and authors differs vastly, and that *c*) this does not depend on the complexity of the spreadsheet at hand [KKG15]. In the latter study we could confirm previously discovered context dimensions of spreadsheets and refine them according to user role and complexity.

In this paper, we want to deepen our understanding of spreadsheets with respect to context by analysing the co-occurrence of context dimensions depending on user role and complexity. As we will use the same data as before and build on those results, we will first shortly introduce the original study in Section 2. Then we proceed with a description of our co-occurrence analysis in Section 3 and discuss potential hypotheses drawn from the results. Section 4 concludes the paper.

2 The Context of Spreadsheet Readers and Authors

The context of a document comprises all explicit and implicit information it contains. From the perspective of human-spreadsheet interaction we are mainly interested in the latter. To get a better understanding of this implicit context, we asked spreadsheet users to explain spreadsheets. On the one hand, we distinguished between spreadsheet authors and readers, as we wanted to learn whether they frame the contained information along different context dimensions. On the other hand, we presented our interview subjects with a complex and a simple spreadsheet. Here, we were interested in understanding the influence of spreadsheet complexity on users' context understanding.

For our study we selected one complex and one simple spreadsheet, both of which are in use at our university and which are manually updated on a regular basis by several employees. ANA GUSEVA in [Gus13] conducted interviews with three spreadsheet readers and three spreadsheet authors of these spreadsheets. The interviews were transcribed and qualitatively analysed with the card-sorting method to classify the elicited data. Here, the transcription is split up into "**knowledge items**", that is, smallest units of information, which are noted down on literal cards. These cards are then sorted into categories. The latter emerge based on the elicited data, they are not given beforehand. These categories provided us with "context dimensions" as they describe how the inter-

viewees framed the information in the given spreadsheets. Concretely, we obtained 319 knowledge items in total and a set of 12 distinct categories; see Table 1.

Dimension	Question
STATEMENT	*What is it? (read keyword)*
REPHRASING	*What is it? (rephrase keyword)*
DEFINITION	*What is it? (formal definition)*
BY-EXAMPLE	*Example?*
EVALUATION	*Is it good?*
FORMULA	*How is it calculated? (function)*
PROVENANCE	*Where does it originate from?*
	How is it calculated? (dependency)
HISTORY	*Has it changed?*
ORGANIZATION	*Where is it?*
PURPOSE	*For what do we need it?*
SIGNIFICANCE	*Is it important?*
OTHER	*Unclassifiable (including false information)*

Table 1. Dimensions and Corresponding Questions after [KKG15] – 'it' refers to the information in the resp. knowledge item

For our context, it is important to note, that a single card, if necessary, was placed under one or more dimensions. The amount of knowledge items placed under a category shows the use rate of the category. We interpret this as its relevance. The exact settings of the study and methods used are described in [KKG15].

The analysis of the categories according to the distinct complexity levels showed a surprising similarity of application and relevance of context dimensions for readers and authors alike. We concluded that complexity makes no difference for the context of readers and authors.

In contrast, there was a big gap in the use of context dimensions by authors and readers when analyzing the categories according to user roles. Here, the difference of the use rates of dimensions EVALUATION, PURPOSE, and SIGNIFICANCE is especially notable. They were vastly underrepresented in the knowledge items of readers. We concluded, that readers need help in this respect.

Sometimes the presentation of numbers and formulae have the effect that people believe that they somehow, implicitly provide the means to assess them, to understand what the author intended with sharing them, and to know what they imply. Thus, if spreadsheets are used as communication tools, these context dimensions need to be acknowledged and their information content has to be made available to readers. To get a deeper understanding how such assistance tools can be of use for readers, we scrutinized the available data material once more. So, in this paper, we study the context dimensions with respect to each other, that is, we analyse the co-occurrences of context dimensions.

3 The Co-Occurrence Study

In this study we distinguish specific sets of knowledge items elicited in [KKG15] according to their occurrence in an interview with a reader or an author and with respect to their referencing the simple or the complex spreadsheet. In particular, we have the following data sets:

All = all available knowledge items ($\# = 319$)
Readers Simple = all knowledge items produced by readers wrt. the simple spreadsheet
Readers Complex = all knowledge items produced by readers wrt. the complex spreadsheet
Authors Simple = all knowledge items produced by authors wrt. the simple spreadsheet
Authors Complex = all knowledge items produced by authors wrt. the complex spreadsheet

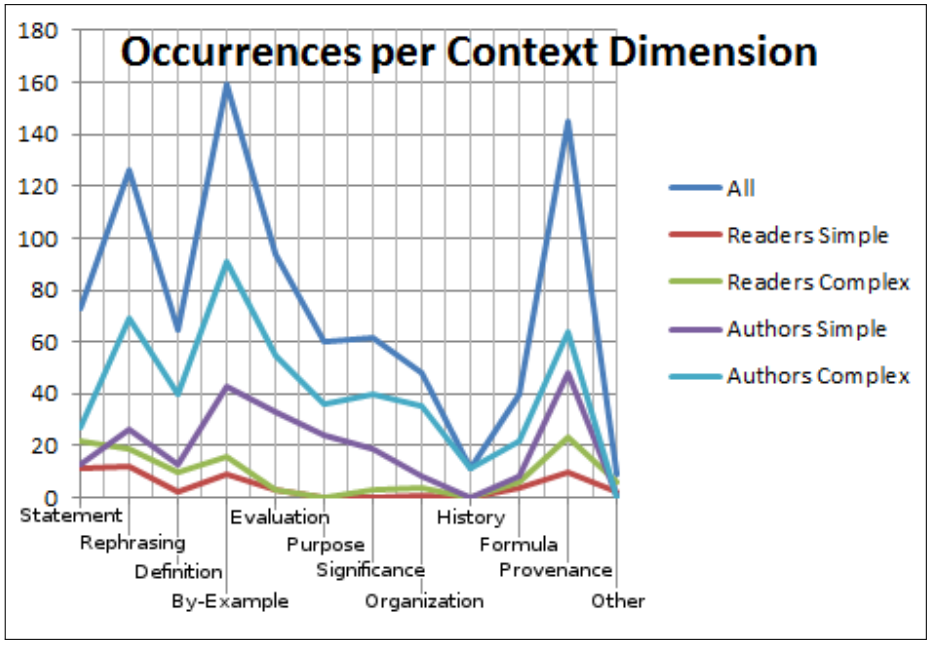

Fig. 1. Number of Occurrences per Context Dimension (x-Axis) and Data Set (y-Axis)

At first, we compare the total occurrences of all knowledge items as a function of context dimensions and data sets (see Figure 1). The occurrences are interesting in the context of studying the co-occurrences as they provide us with a sense of relevance of context dimensions with respect to the specific data sets. Therefore, the former allow us a better interpretation of the latter. We immediately observe:

– It is obvious that readers used fewer explanations than authors, i.e. they said less about the spreadsheets at hand. The probable reason is our result in [KKG15] and we include it here for completeness reasons:

  Hypothesis 1 ("Readers need Help"):
  "*Readers don't know as much as authors about the spreadsheet context.*"

– Within this pattern we recognize that

  Hypothesis 2 ("Complexity and Context"):
  "*The more complex the spreadsheet the more the users put the context into words.*"

  Note that this is equally true for readers as for authors. The more complex the spreadsheet the more users try to find the right frame to put the information into, so that they grasp the information with its full implications.

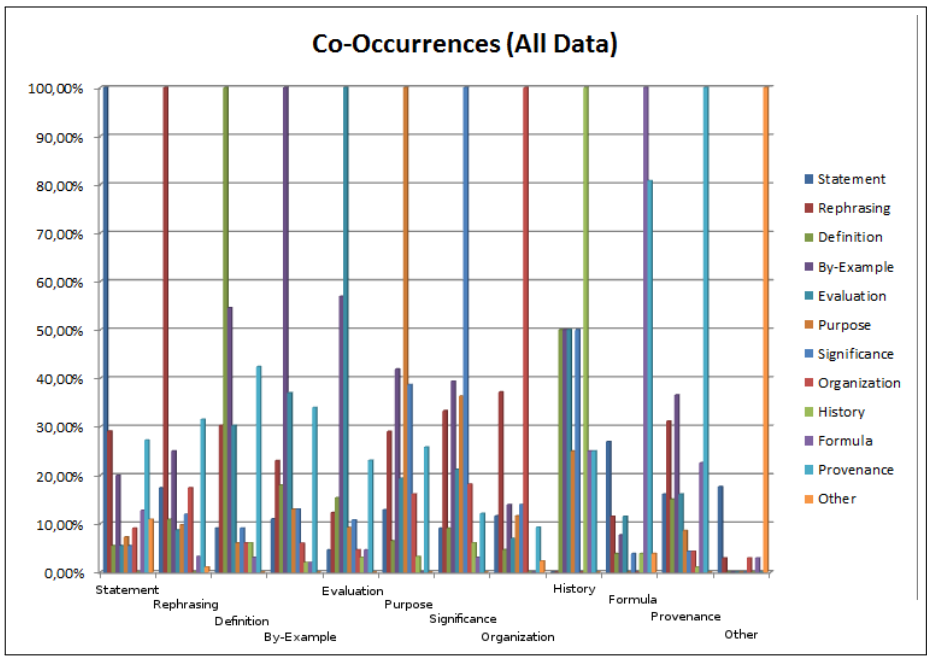

Fig. 2. The Overall Co-Occurrence Probabilities

– There is a very striking minimum of knowledge items by readers for the context dimensions EVALUATION, PURPOSE, and SIGNIFICANCE. This interesting result was elaborated on in [KKG15]. Again, we repeat it for completeness reasons:

  Hypothesis 3 ("From Information to Knowledge"):
  "*Readers don't grasp the full set of implications of spreadsheet data.*"

  According to [PRR97] The difference between "data" and "information" consists in the provision of local context, whereas the difference between "information" and "knowledge" consists in the provision of global context [ibd]. The global context

allows humans to predict the implications of their actions, so it is essential for planning their future actions based on what they perceive.

– As the simple spreadsheet really had no history and the history of the complex one was unknown to the readers, the global minimum for HISTORY was to be expected.

Next, we compute co-occurrences $c_{X,Y}(k)$, where the "**co-occurrence of the context dimensions X and Y** $c_{X,Y}(k)$ for a knowledge item k" is 1 if the knowledge item was sorted into context dimension X and Y and 0 elsewhere. The sum over all n knowledge items k from a data set for given X and Y gives us a non-knowledge-item-dependent measure $c_{X,Y}$, that is the higher the more often the co-occurrence over the set of knowledge items happened:

$$c_{X,Y} = \sum_{k=1}^{n} c_{X,Y}(k)$$

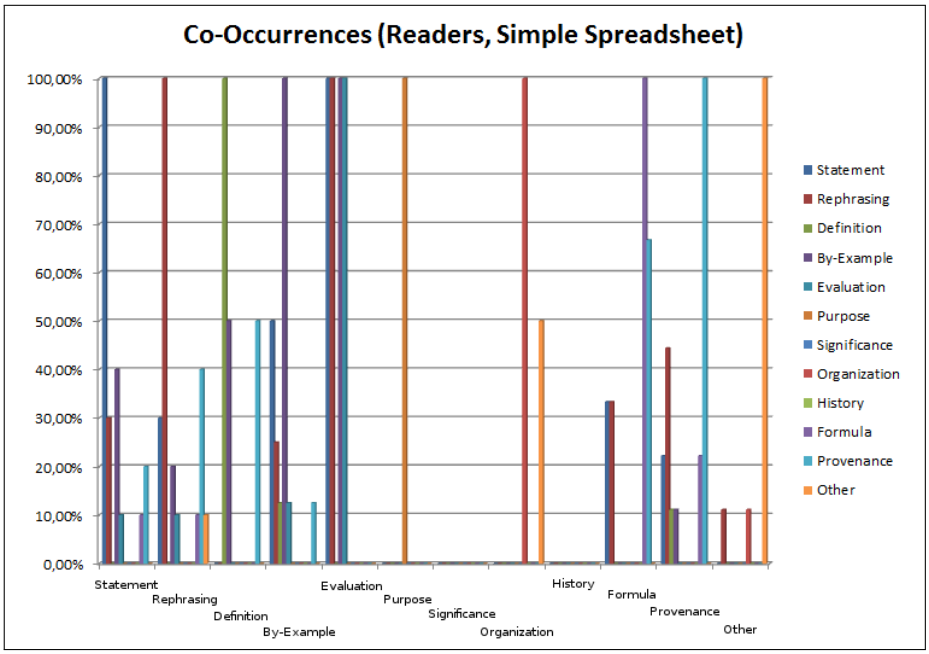

Fig. 3. Readers' Co-Occurrence Probabilities for the Simple Spreadsheet

Therefore, we define $P(Y|X)$ as the "**co-occurrence probability**" for any knowledge item sorted into dimension X to be also sorted into dimension Y, i.e.,

$$P(Y|X) = \begin{cases} \frac{c_{X,Y}}{c_{X,X}} & \text{for all } X \text{ in which } c_{X,X} > 0 \\ 0 & elsewhere \end{cases} \quad (1)$$

The matrix filled with components $P(Y|X)$ yields a distribution of the co-occurrence probabilities, which can be seen in Figure 2 for the set of all 319 knowledge items. In particular here, we can see twelve blocks B_X. In each B_X it holds that $P(X|X) = 100\%$ for the given X. Moreover, in each B_X the other $P(Y|X)$, that is, the co-occurrences of these dimensions with dimension X, are visualized.

We observe, that the context dimension STATEMENT has low co-occurrence probabilities with respect to the other context dimensions, but, e.g., the co-occurrence between FORMULA and PROVENANCE is rather high.

We acknowledge that the danger of interpretation lies in disregarding the difference of correlation and causality. Therefore we only consider this paper as a pre-study, in which hypotheses are elicited that have to be proven elsewhere. But we can already hypothesize from the data that

Hypothesis 4 ("Documenting Formulae"):
"*Formulae provide spreadsheet users with a security of provenance of data.*"

Analogously, consider the other data sets visualized below.

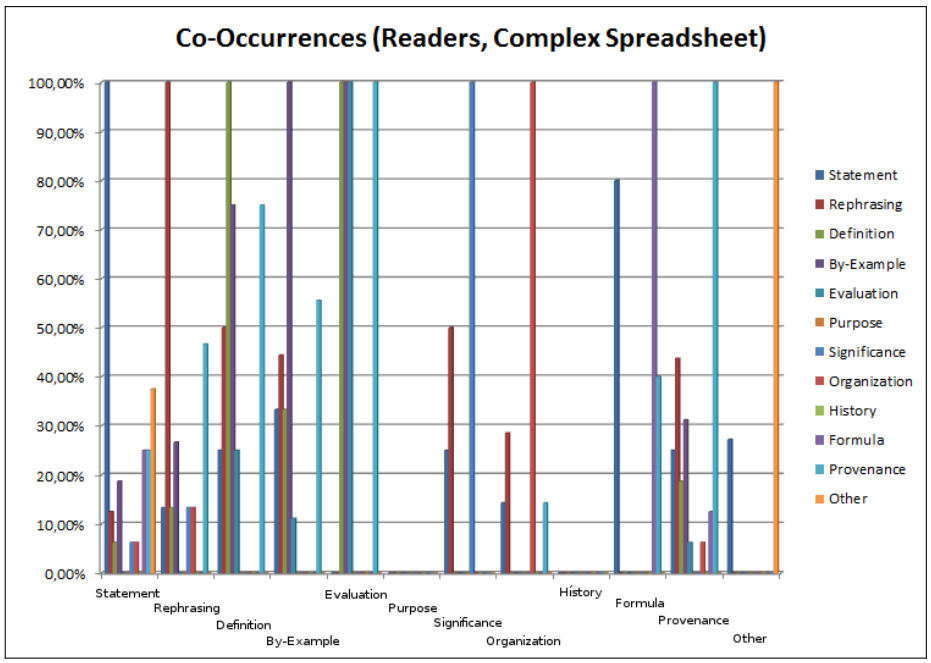

Fig. 4. Readers' Co-Occurrence Probabilities for the Complex Spreadsheet

In Figure 3 we see the co-occurrence distribution of readers commenting the simple spreadsheet. Here, it is remarkable that there are almost identical co-occurrence probabilities for a given EVALUATION with other context dimensions, specifically STATEMENT, REPHRASING, and BY-EXAMPLE, but no others at all. A hypothesis could be that if a reader is ready to give an evaluation, then she is not only also able to give a statement about the content, she can moreover rephrase it and give an example.

Hypothesis 5 ("Evaluation Hinge"):
"*Evaluation depends on the ability of providing an example.*"

We might even draw from this the conjecture that the provision of STATEMENT, REPHRASING, and BY-EXAMPLE are necessary, if we want the reader to be able to evaluate the data.

If we didn't look at Figure 1, we could interpret into Figure 3 that a reader of a simple spreadsheet finds an explanation of type PURPOSE sufficient as no other co-occurrence appeared. But as there are almost no explanations of this type (see Figure 1), this statement wouldn't make any sense at all.

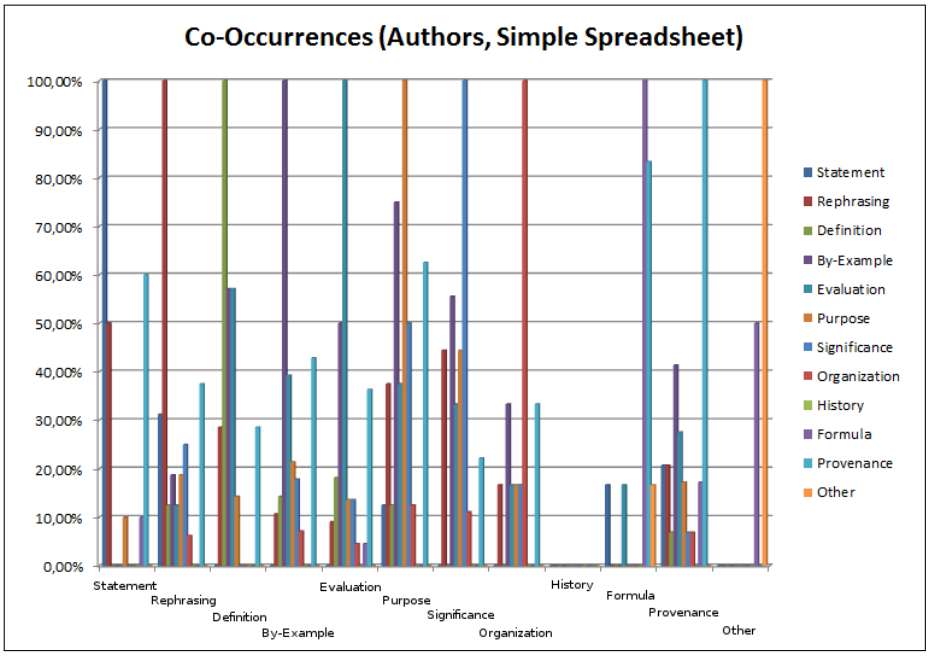

Fig. 5. Authors' Co-Occurrence Probabilities for the Simple Spreadsheet

The diagram for the co-occurrence distribution of context dimensions for readers of a complex spreadsheet seen in Figure 4 is equally empty as the one in Figure 3. Here, we note that the dimension ORGANIZATION is suspiciously missing or very low. For complex spreadsheets we had conjectured before that readers might make use of organization to understand the content of spreadsheet, but they didn't try very much.

We also looked at the data for the authors of simple and complex spreadsheets as can be seen in Figures 5 and 6. We directly observe that the co-occurrence probabilities tend to be higher than with the readers context dimensions. As was the case for the diagrams for the readers, the co-occurrence probability of a given FORMULA knowledge item being also in the PROVENANCE dimension is very high. In general the authors seem to have been able to make use of almost all other context dimensions to illustrate their comprehension of a spreadsheet.

Note that authors did not use any description for the HISTORY dimension of spreadsheet context in the simple case – maybe there was no history. In contrast, only the

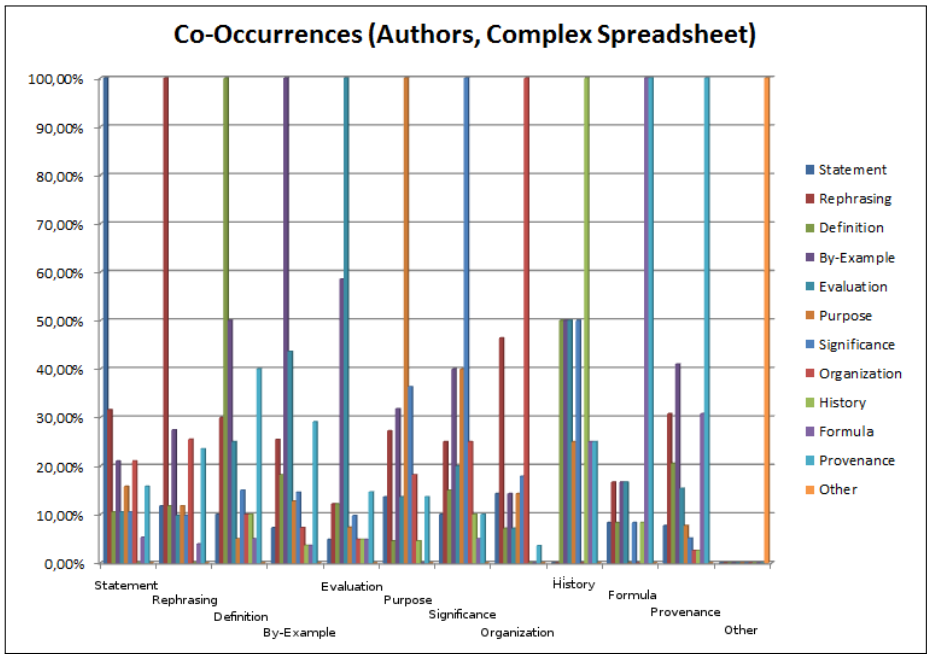

Fig. 6. Authors' Co-Occurrence Probabilities for the Complex Spreadsheet

authors tried to explain the history of the complex spreadsheet and it seemed to be worthwhile for them to describe this within several context dimensions.

4 Conclusion

In this paper, we had a first look at co-occurrences of context dimensions of spreadsheet users. The observations from this data can be used as a pre-study which set up hypotheses with respect to the relationship between distinct context-dimensions. In summary, the following hypotheses were inferred:

Hypothesis 1: "*Readers don't know as much as authors about the spreadsheet context.*"
Hypothesis 2: "*The more complex the spreadsheet the more the users put the context into words.*"
Hypothesis 3: "*Readers don't grasp the full set of implications of spreadsheet data.*"
Hypothesis 4: "*Formulae provide spreadsheet users with a security of provenance of data.*"
Hypothesis 5: "*Evaluation depends on the ability of providing an example.*"

If we can confirm or reject in future studies some of the hypotheses yielded by this research, then we can take those as design suggestions for future user assistance systems. We believe that these results are not only limited to human-spreadsheet interaction, but can be generalized to mathematical user-interfaces in general as long as they are used for communication purposes in a broad sense.

In hindsight, the definition of co-occurrence should have also included the reference to the concrete spreadsheet content as the reasoning with respect to the concrete content had to be left out in this study, but would have been interesting.

Acknowledgement We are very grateful for the valuable comments given by the anonymous reviewers of the paper and thank them for their careful and pointed reviews.

References

[Bak+08] Kenneth R. Baker et al. "Comparison of Characteristics and Practices amongst Spreadsheet Users with Different Levels of Experience". In: *CoRR* abs/0803.0168 (2008).

[CMW07] Jonathan P. Caulkins, Erica Layne Morrison, and Timothy Weidemann. "Spreadsheet Errors and Decision Making: Evidence from Field Interviews". In: *JOEUC* 19.3 (2007), pp. 1–23.

[CS10] Chris Chambers and Chris Scaffidi. "Struggling to Excel: A Field Study of Challenges Faced by Spreadsheet Users". In: *Proceedings of the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing*. VLHCC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 187–194. ISBN: 978-0-7695-4206-5.

[Gus13] Ana Guseva. *Towards Understanding Context Dimensions of Spreadsheet Knowledge*. 2013.

[HG93] David G. Hendry and Thomas R. G. Green. "CogMap: a Visual Description Language for Spreadsheets". In: *J. Vis. Lang. Comput.* 4.1 (1993), pp. 35–54.

[HG94] David G. Hendry and Thomas R. G. Green. "Creating, comprehending and explaining spreadsheets: a cognitive interpretation of what discretionary users think of the spreadsheet model". In: *Int. J. Hum.-Comput. Stud.* 40.6 (1994), pp. 1033–1065.

[KKG15] Andrea Kohlhase, Michael Kohlhase, and Ana Guseva. "Context in Spreadsheet Comprehension". In: *Second workshop on Software Engineering methods in Spreadsheets*. accepted. 2015. URL: `http://kwarc.info/kohlhase/submit/sems15-context.pdf`.

[Koh10] Andrea Kohlhase. "Towards User Assistance for Documents via Interactional Semantic Technology". In: *KI 2010: Advances in Artificial Intelligence*. Ed. by Rüdiger Dillmann et al. LNAI 6359. Karlsruhe, Germany, 2010, pp. 107–115.

[Koh13] Andrea Kohlhase. "Human-Spreadsheet Interaction". In: *Human-Computer Interaction – INTERACT 2013*. Ed. by Paula Kotzé et al. LNCS 8120. Heidelberg: Springer, 2013, pp. 571–578. ISBN: 978-3-642-40497-9.

[NM90a] Bonnie A. Nardi and James R. Miller. "An Ethnographic Study of Distributed Problem Solving in Spreadsheet Development". In: *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*. ACM Press, 1990, pp. 197–208.

[NM90b] Bonnie A. Nardi and James R. Miller. "The spreadsheet interface: A basis for end user programming". In: *Proceedings of the IFIP TC13 Third Interational Conference on Human-Computer Interaction*. INTERACT '90. Amsterdam, The Netherlands, The Netherlands: North-Holland Publishing Co., 1990, pp. 977–983. ISBN: 0-444-88817-9.

[Pan00] Raymond R. Panko. “Spreadsheet Errors: What We Know. What We Think We Can Do.” In: *Symp. of the European Spreadsheet Risks Interest Group (EuSpRIG 2000)*. 2000.

[PBL08] Stephen G. Powell, Kenneth R. Baker, and Barry Lawson. “A critical review of the literature on spreadsheet errors”. In: *Decision Support Systems* 46.1 (2008), pp. 128–138.

[PLB08] Stephen G. Powell, Barry Lawson, and Kenneth R. Baker. “Impact of Errors in Operational Spreadsheets”. In: *CoRR* abs/0801.0715 (2008).

[PRR97] G. Probst, St. Raub, and Kai Romhardt. *Wissen managen*. 4 (2003). Gabler Verlag, 1997.

[SSM05] Christopher Scaffidi, Mary Shaw, and Brad A. Myers. “Estimating the Numbers of End Users and End User Programmers”. In: *VL/HCC*. 2005, pp. 207–214.

[Wol+11] K. Wolstencroft et al. “RightField: Embedding ontology annotation in spreadsheets”. In: *Bioinformatics* 24.14 (2011), pp. 2021–2022.