Supplementary material

The following content represents material supplementary to the PhD thesis "Towards automatically generating supply chain maps from natural language text" by Pascal Wichmann. The content may provide additional valuable insights but is not at the core of the PhD work. The latest version of this document should be accessible via the following URL: `https://github.com/pwichmann/phd_thesis`. The content in the supplementary material covers the following topics:

- Distinction between supply chain and related concepts
- Supply chain resilience
- A critical reflection on the value of supply chain maps
- Inter- and intra-annotator agreement
- Use cases of structural supply chain visibility

Distinction between supply chain and related concepts

Distinction between supply chain and value chain

The terms *supply chain* and *value chain* sound similar and can easily be confused. The concept of a "value chain" was introduced by Michael Porter in the context of a firm's competitive advantage. "Competitive advantage cannot be understood by looking at a firm as a whole. It stems from the many discrete activities a firm performs in designing, producing, marketing, delivering, and supporting its product. Each of these activities can contribute to a firm's relative cost position and create a basis for differentiation. [...] A systematic way of examining all the activities a firm performs and how they interact is necessary for analyzing the sources of competitive advantage. [The value chain is] the basic tool for doing so. The value chain disaggregates a firm into its strategically relevant activities in order to understand the behaviour of costs and the existing and potential sources of differentiation. A firm gains competitive advantage by performing these strategically important activities more cheaply or better than its competitors." (Porter, 1985, p. 33). The concept of a value chain distinguishes *primary activities*[1] and *support activities*[2]. Unlike the concept of a supply chain,

[1] Inbound logistics, Operations, Outbound logistics, Marketing and sales, Service

[2] Procurement, Technology development, Human resource management, and Firm infrastructure

the value chain framework captures all the activities performed *within a firm*. Nevertheless, there is a connection between supply chain and value chain that has been highlighted by (Christopher, 2005). Porter's thesis is that firms should analyse their activities and consider outsourcing those ones which cannot be performed with a competitive advantage. "The effect of outsourcing is to extend the value chain beyond the boundaries of the business. In other words, the supply chain becomes the value chain. Value (and cost) is created not just by the focal firm in a network, but by all the entities that connect to each other." (Christopher, 2005, p. 14).

Distinction between supply chain and value stream

A further term shall be briefly discussed to help distinguish different mapping techniques at a later point in this chapter. In the context of *lean manufacturing*, the term *"value stream"* is used to describe a concept that is similar to the concept of a supply chain. According to Rother and Shook (1999), a value stream is "is all the actions (both value added and non-value added) currently required to bring a product through the main flows essential to every product: (1) the production flow from raw material into the arms of the customer, and (2) the design flow from concept to launch".

Distinction between supply chain mapping and value stream mapping

In the context of *lean manufacturing* and *business process re-engineering*, the term *value stream mapping*[3] is used to describe a specific technique. Rother and Shook (1999) provide the following definition: "Value stream mapping is a pencil and paper tool that helps you to see and understand the flow of material and information as a product makes its way through the value stream. What we mean by value stream mapping is simple: Follow a product's production path from customer to supplier, and carefully draw a visual representation of every process in the material and information flow. Then ask a set of key questions and draw a 'future state' map of how value should flow.". The lean management concept has since also been applied to supply chain management (as "lean supply chain management") and methods like "supply chain value stream mapping" (see Suarez-Barraza et al. (2016)) have been proposed which blur the lines between supply chain mapping and value stream mapping. Generally, value stream mapping is process-based and seeks to identify and eliminate waste along the value stream. The maps often use icons, such as factory icons, and data boxes to hold information about requirements, such as the required number of products per month or

[3]Within the Toyota Production System, this technique was called "material and information flow mapping".

on-time delivery information. Figure 1 provides an example of a supply chain value stream map by Suarez-Barraza et al. (2016).

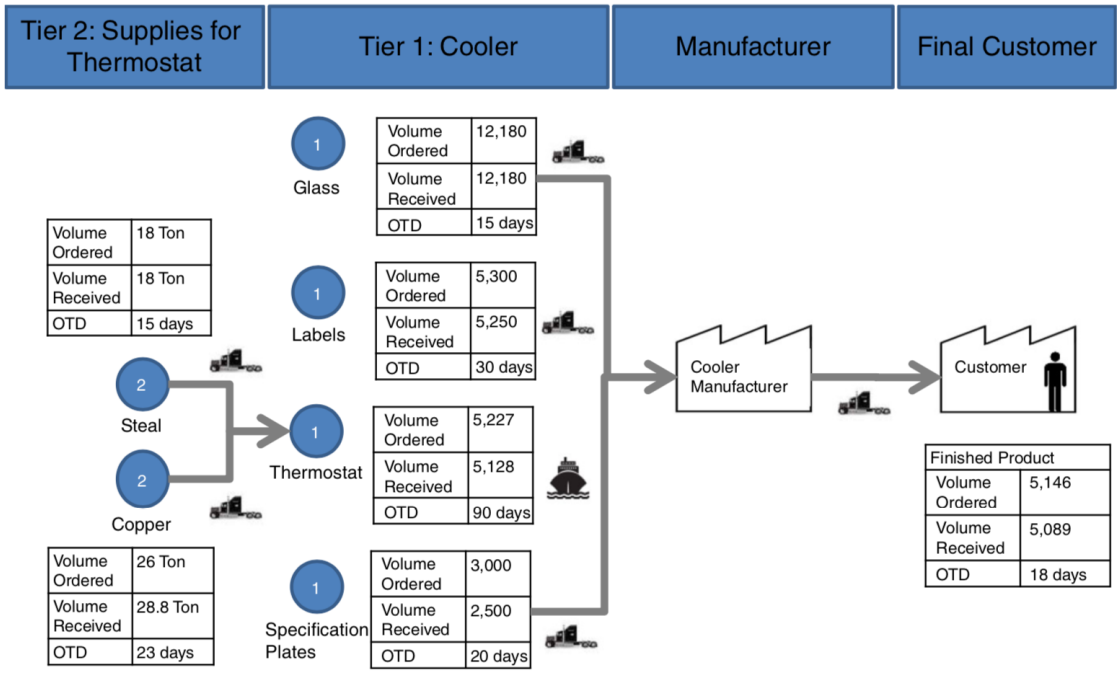

Fig. 1 Supply chain value stream map (here: showing the material flow only), from Suarez-Barraza et al. (2016)

Further work on applying value stream mapping to multiple tiers of a supply chain (as opposed to processes within a single firm) can be found, for instance, in Boonsthonsatit and Jungthawan (2015) and Anderson (2017). Anderson (2017) also provides a comparison of value stream mapping and the supply chain reference model SCOR[4].

[4]The *Supply Chain Operations Reference* (SCOR) is a "supply chain reference model" or "supply chain process framework". Originally, SCOR was developed in 1996 by PRTM, a management consulting firm, and was endorsed by the Supply-Chain Council which is now a part of APICS. APICS has been maintaining the SCOR model (`http://www.apics.org/apics-for-business/frameworks/scor`; last accessed: 2019-08-01). The most recent version of the framework, SCOR 12.0, was released in 2017 by APICS (`http://www.apics.org/apics-for-business/frameworks/scor12`; last accessed 2019-08-01). The model aims at benchmarking supply chain performance and tracking improvements over time. SCOR organises all the processes in a supply chain into six groups: Plan, Source, Make, Deliver, Return, and Enable.

Supply chain resilience

A comprehensive overview of supply chain resilience definitions can be found in Kim et al. (2015) as well as in Pettit (2008).

Christopher and Peck (2004) define supply chain resilience as "the ability of a system to return to its original state or move to a new, more desirable state after being disturbed" (Christopher and Peck, 2004). In their explanations, the authors explicitly distinguish resilience from robustness – unfortunately, they only do so based on dictionary definitions of both terms instead of on the basis of academic or more context-specific robustness definitions, and they do not clarify the inter-relation of both concepts. Christopher and Peck stress the notion of flexibility, adaptability (in the sense of moving to a new state) or agility (in the sense of quickly reacting to unpredictable events).

Ponomarov and Holcomb (2009) provide a more specific description of supply chain resilience by defining it as "the adaptive capability of the supply chain to prepare for unexpected events, respond to disruptions, and recover from them by maintaining continuity of operations at the desired level of connectedness and control over structure and function" (Ponomarov and Holcomb, 2009). This definition explicitly accounts for different phases in the process of managing risks ("readiness", "responsiveness" and "recovery") that had been proposed by Sheffi (2005). Jüttner and Maklan (2011) adopt the definition by Ponomarov and Holcomb and propose four formative capabilities for supply chain resilience: (1) flexibility, (2) velocity, (3) visibility and (4) collaboration (Jüttner and Maklan, 2011). They also define "agility" as the combination of visibility, velocity and flexibility.

As with the general term resilience, both notions of stability (topology/structure) and flexibility (dynamics) can be considered part of supply chain resilience.

Table 1 Overview of selected *supply chain resilience* definitions

Reference	Definition
Christopher and Peck (2004)	"the ability of a system to return to its original state or move to a new, more desirable state after being disturbed"
Sheffi (2005)	"the ability to bounce back from a disruption [...] Resilience [...] can be achieved by either creating redundancy or increasing flexibility."
Ponomarov and Holcomb (2009)	"the adaptive capability of the supply chain to prepare for unexpected events, respond to disruptions, and recover from them by maintaining continuity of operations at the desired level of connectedness and control over structure and function"
Jüttner and Maklan (2011)	No explicit new definition of supply chain resilience provided, but definition via formative elements flexibility, velocity, visibility, and collaboration (based on Ponomarov and Holcomb (2009))

Jüttner and Maklan propose an inter-relation between the three concepts supply chain risk, supply chain resilience and supply chain vulnerability – as shown in Figure 2.

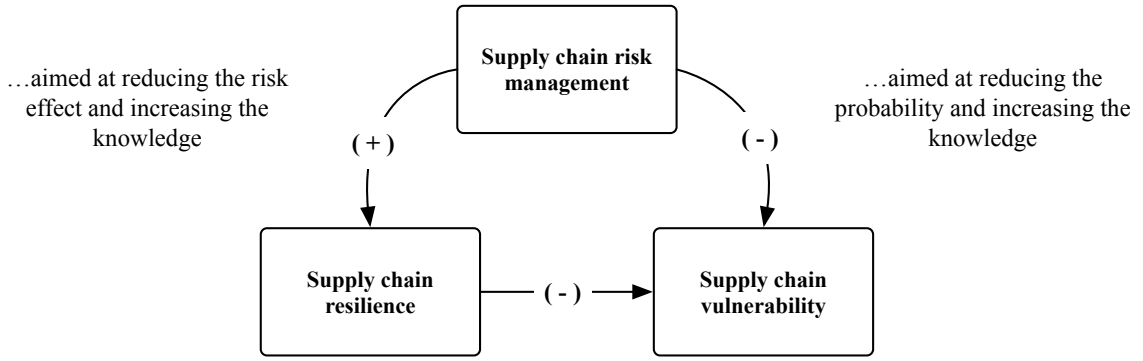

Fig. 2 The relationship between supply chain resilience, supply chain risk management and supply chain vulnerability; source: author; adapted from Jüttner and Maklan (2011)

Some authors, such as Pettit (2008), see the concept of supply chain resilience extend traditional risk management approaches by also addressing unforeseeable events, whereas traditional risk management starts with the step of risk identification.

Another finding by Christopher and Peck is noteworthy: "There are certain features that, if engineered into a supply chain, can improve its resilience" (Christopher and Peck, 2004). Hence, supply chains can be designed or actively be re-engineered to become more resilient.

Within the scope of this research, we will adopt a definition of supply chain resilience that is compatible with the objective of general (supply chain) risk management. Hence, the spectrum of resilience-related actions encompasses both mitigation and contingency. It includes *proactive* measures to reduce the likelihood of disruption, reduce the possible impact should a disruption occur, and increase the lead time of detection as well as *reactive* measures to contain the impact once a disruption has occurred, i.e. the ability to quickly respond and recover. A definition of resilience only seems sensible if it includes the ability to recover swiftly but also to prevent disruptions in the first place.

Similar to the general concept of resilience, supply chain resilience conflicts with other objectives of supply chain management. While redundant inventory would increase resilience, it would counteract the ideals of efficiency and lean management and increase operating costs.

A critical reflection on the value of supply chain maps

The theoretical and practical evidence in support of the value of supply chain maps has been provided in Chapter 2 as well as Chapter 4. This section aims to reason about this more critically, especially from the standpoint of supply chain maps of *imperfect* information quality.

Imperfect information

Due to the imperfect quality (including availability) of the input data, independent of the method, any practical attempt to create multi-tiered supply chain maps is likely to result in maps that are *not fully complete*, *not perfectly accurate*, or *uncertain* (probabilistic). With such expectations, one could be inclined to dismiss *any* attempts to map the supply chain.

One could take the position that an imperfect supply chain map may obscure a (major) vulnerability and hence provide a false sense of security. And because a single bad sub-tier supplier can change a company's risk profile, a map not including exactly that sub-tier supplier will give a misleading impression. Furthermore, one could argue that *inaccurate* information may not only be useless but harmful if costly actions are triggered based on what turns out to be a false positive. *Incomplete* information could be harmful if other measures of trying to understand the supply chain are neglected.

On the other hand, a number of arguments can be presented to support the effort of extracting supply chain maps in spite of imperfect information quality:

Information about supply chain structures *is* valuable The literature review has shown unequivocally that information about supply chain structures is relevant for a wide range of managing tasks with the potential to change decisions. For instance, supply chain vulnerabilities may not be visible in historical delivery data.

Information about supply chain structures *can* be valuable *even if the information is imperfect* Even if information about a company's supply chain is not fully complete or not fully accurate, this information can be valuable and is not necessarily worthless. Instead, there will be some value continuum specific to the use case at hand. It is immediately evident that incomplete information *can* be valuable in at least some use cases: Detecting a single, previously unknown, undesirable sub-tier supplier is actionable information for a company – without the need to know all of the remaining supply chain across all tiers. Even if the overall risk exposure cannot be assessed, at least specific vulnerabilities can be identified and addressed. Lastly, incremental improvements matter. A small increase in information, more accurate information, more certainty will always be better than not having this improvement (if information acquisition costs are acceptable).

No method can provide *perfect* supply chain maps It appears that there is no other method (supplier surveys, commercial third-party databases, ...) that would provide *perfect* supply chain maps. Furthermore, it seems that automated mapping from text documents is at least a scalable, cost-efficient *complementary* approach. If supply chain maps were only valuable if they were fully complete and perfectly accurate, then the currently available datasets should not have any commercial value; yet they sell for substantial amounts of money.

Aiming for a 100% complete map is unrealistic It is unlikely that supply chains are fully described by publicly available text documents. And, thus, the vision for this research is not to reconstruct 100% of any supply chain but at least to extract all the information that is available. This extracted information can then be used to complement, rather than replace, other approaches (such as supplier surveys). An automated approach could also be used to crosscheck or regularly update existing supply chain maps as well as to quickly generate a rudimentary overview.

Some actions are cost-efficient – even if taken erroneously Some actions triggered by information obtained from a supply chain map can be effective while also being cost-efficient – even if the action was taken based on false information. Such actions could simply be aimed

at acquiring more information. Giving a supplier a call to obtain more information or to hint at a problematic sub-tier supplier does not incur major costs.

Further aspects

Specialised approaches will exceed general-purpose approaches in performance One could argue that OpenIE systems will soon be able to convert any text into structured data for any relation. Open IE or general ontology learning are likely to further improve over time and also be able to extract basic company-to-company relations. However, specialised approached will exceed their performance for the domain they have been designed for (since specialised systems can just integrate general solution and build on top). And, thus, it is not unreasonable to work on a specialised approach.

Plausible deniability may not be a viable long-term strategy In one of the interviews with supply chain managers, it was stated that the legal department of that company would prefer them to *not* try to better understand the extended supply chain because this might increase the company's liability. From a legal standpoint, it would be safer not to know. This appears to be at least ethically questionable and may not be a viable long-term strategy.

Establishing the correct tier of a company Related to the problem of information quality is the problem of not having sufficient or sufficiently unambiguous information to correctly establish the tier a company is operating on for a particular supply chain. The problem and potential solutions are discussed in more detail in Section **??**.

Increased structural visibility may not only have beneficiaries It is plausible that an increase in structural supply chain visibility may harm some players. This includes companies with unsustainable business practices but it can also include sub-tier suppliers that may now be under tighter control by customers further downstream the supply chain.

Inter- and intra-annotator agreement

Inter-annotator agreement

Importance of inter-annotator agreement

Measuring inter-annotator agreement (also known under different names, such as "inter-rater reliability") is a crucial part of the annotation collection. As common in Natural Language

Processing and Machine Learning when manual annotations have to be collected, one is interested in the *correctness or validity* of these labels. But since linguistic categories are determined by human judgement, correctness cannot be measured directly. And so, instead of measuring correctness, one measures the *reliability* of annotation. If human annotators *consistently* make same decisions, so the assumption, then they have internalised the annotation scheme and, thus, high reliability suggests validity of the collected labels. Because correctness ("ground truth") cannot directly be measured but only approximated by the measured consensus, a high inter-annotator agreement does not necessarily mean that the annotations are correct. The result can be influenced by a number of factors and inter-annotator agreement becomes an indicator of the inherent difficulty of the task as well as the quality of the annotations. If annotators *cannot* agree on the correct answer, the task may be too difficult, the task instructions too unclear, the results not reproducible, and a successful automation of the task may not be possible.

Metrics of inter-annotator agreement

A wide range of metrics for inter-annotator agreement have been designed, such as Cohen's Kappa (κ) (Cohen, 1960), Fleiss' Kappa (Fleiss, 1971), Scott's Pi (π) (Hallgren, 2012; Scott, 1955) or Krippendorff's Alpha (α) (e.g. see Krippendorff (2011)). There is still an ongoing debate in the scientific community whether or not these metrics appropriately measure all aspects of agreement in an annotation task (see Pustejovsky and Stubbs (2013, p. 126) who also provide further references on the debate).

Even though metrics of inter-annotator agreement are useful and commonly used, a few limitations have to be kept in mind: (1) The metrics measure consensus, not ground truth. (2) There is no universally accepted interpretation of the metric scales, e.g. an agreement of 0.8 can be poor or exceptionally good depending on the type of task.

Cohen's Kappa and Fleiss' Kappa shall be summarised in more detail in the following paragraphs.

Cohen's Kappa

Cohen's Kappa (κ) measures the agreement between *two* annotators (having annotated the exact same set of examples) while taking into account the possibility of agreement by chance (Pustejovsky and Stubbs, 2013). Agreement by chance in this context does not just naively assume equal probabilities for all categories and annotators but takes into account the probability that a specific annotator chooses a specific class – based on the observed annotation choices. Cohen's Kappa is defined by the following equation:

$$\text{Cohen's } \kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{1}$$

In the equation above, p_o stands for the observed agreement between the annotators and p_e stands for the expected agreement between annotators, if each annotator was to randomly pick a category for each annotation. Like most correlation statistics, the values for κ can range from -1 to +1, where a value of 0 represents the amount of agreement that can be expected from random chance alone, and a value of 1 represents perfect agreement between the annotators. While negative κ values are theoretically possible, they are unlikely in practice and would indicate an agreement worse than random with -1 indicating "perfect" disagreement. While various scales for the interpretation of κ values between 0 and 1 have been suggested (e.g. see Landis and Koch (1977)), the interpretation cannot be standardised and depends on the task at hand. Cohen's κ assumes an equal weight for all annotators and does not assume any order of the categories. Cohen's κ requires that both annotators provided labels for the exact same set of examples. To adapt the metric from a pair-wise comparison to more than two annotators, the arithmetic average of Cohen's κ across all pairs of annotators can be computed (see Hallgren (2012)). In the following section, a worked example of Cohen's κ is provided.

Example calculation of Cohen's Kappa statistic

This section shall provide an example calculation of Cohen's Kappa statistic. The assumed confusion matrix is provided by Figure 3[5]. The matrix shows for all combinations of available classes how many instances were labelled this way by Annotator 1 and Annotator 2. The diagonal elements highlighted in green contain the number of instances where both annotators agreed on the class. The blue ellipses show the sums over columns or rows, the orange ellipses show the probabilities that an annotator chooses a specific class label (based on the observed labelling decisions).

[5]The values are based on an example in Pustejovsky and Stubbs (2013). Errors found in the reference were corrected.

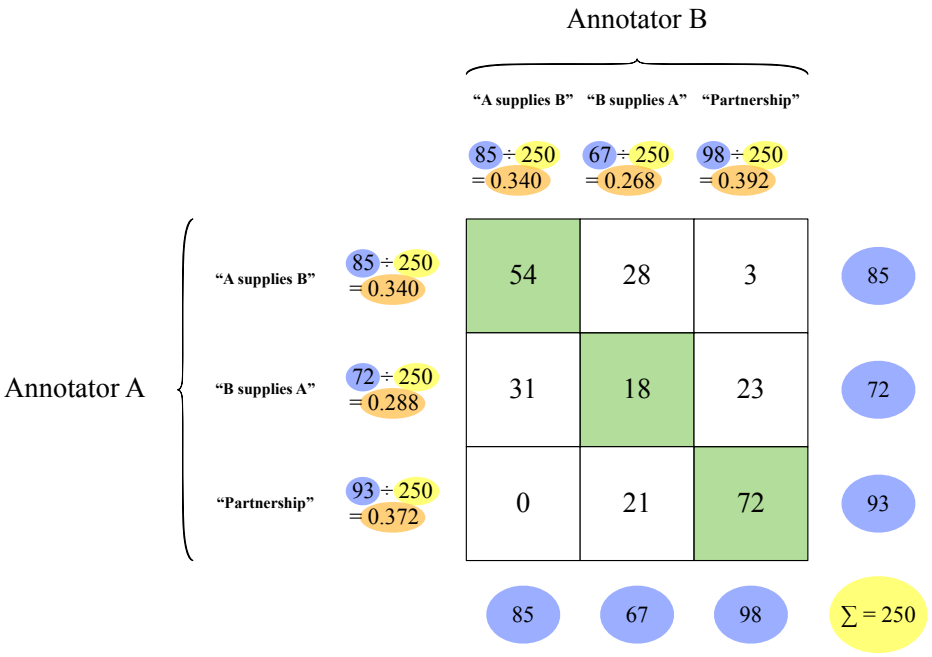

Fig. 3 Confusion matrix for an example calculation of Cohen's Kappa

Cohen's Kappa is defined by the following equation:

$$\text{Cohen's } \kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{2}$$

In the equation above, p_o stands for the observed agreement between the annotators and p_e stands for the expected agreement between annotators, if each annotator was to randomly pick a category for each annotation.

Calculation of the observed agreement (p_o) The observed agreement between the annotators is the sum of all diagonal elements divided by the overall number of instances, in this case:

$$p_o = \frac{54 + 18 + 72}{250} = \frac{144}{250} = 0.576 \tag{3}$$

Calculation of expected agreement between annotators by chance alone (p_e) To determine expected agreement between annotators by chance alone (p_e), one first determine the probabilities that the two annotators would agree on class 1, class 2, and class 3. These probabilities are then summed up to obtain p_e.

① **A supplies B:** Based on the given confusion matrix, the chance of both annotators randomly choosing "A supplies B" is equal to:
$0.34 \cdot 0.34 = 0.1156 \approx 0.116$

② **B supplies A:** Based on the given confusion matrix, the chance of both annotators randomly choosing "B supplies A" is equal to:
$0.288 \cdot 0.268 = 0.077184 \approx 0.077$

③ **Partnership:** Based on the given confusion matrix, the chance of both annotators randomly choosing "Partnership" is equal to:
$0.372 \cdot 0.392 = 0.145824 \approx 0.146$

The expected agreement p_e is the sum of these probabilities:

$$p_e = ① + ② + ③ = 0.116 + 0.077 + 0.146 = 0.339 \tag{4}$$

Calculating Cohen's κ Cohen's κ can now be calculated as:

$$\text{Cohen's } \kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.576 - 0.339}{1 - 0.339} = \frac{0.237}{0.661} = 0.35854 \approx 0.359 \tag{5}$$

Fleiss' Kappa

A common option for measuring consensus across more than just two annotators, is *Fleiss' Kappa* (Fleiss, 1971), which looks similar to Cohen's Kappa but is actually an extension of Scott's Pi (π) statistic (Hallgren, 2012; Scott, 1955). Fleiss' κ assumes an equal weighting of annotators and the order of the categories is assumed to be unimportant. Note that Fleiss' Kappa does *not* assume that all items are annotated by the *same* annotators, but it *does* assume that all items are annotated the same number of times[6]. In the following section, the definition of Fleiss' Kappa and a worked example to calculate Fleiss' κ are provided.

Definition and worked example for Fleiss' Kappa

Data representation and assumptions As opposed to the confusion matrix used for Cohen's Kappa, the annotation is represented differently for the computation of Fleiss' Kappa,

[6]Taken from the errata of (Pustejovsky and Stubbs, 2013) published online

where a matrix of Examples $\times$ Classes is used. This example assumes six annotators, 3 classes, and 10 examples that required classification. Each number in the matrix reflects how often an annotator has assigned a specific class.

While for Cohen's κ both annotators evaluate every example, in the case of Fleiss' κ, there can be many annotators and not every annotator needs to evaluate each example; what is important is that each example is evaluated exactly the identical number times.

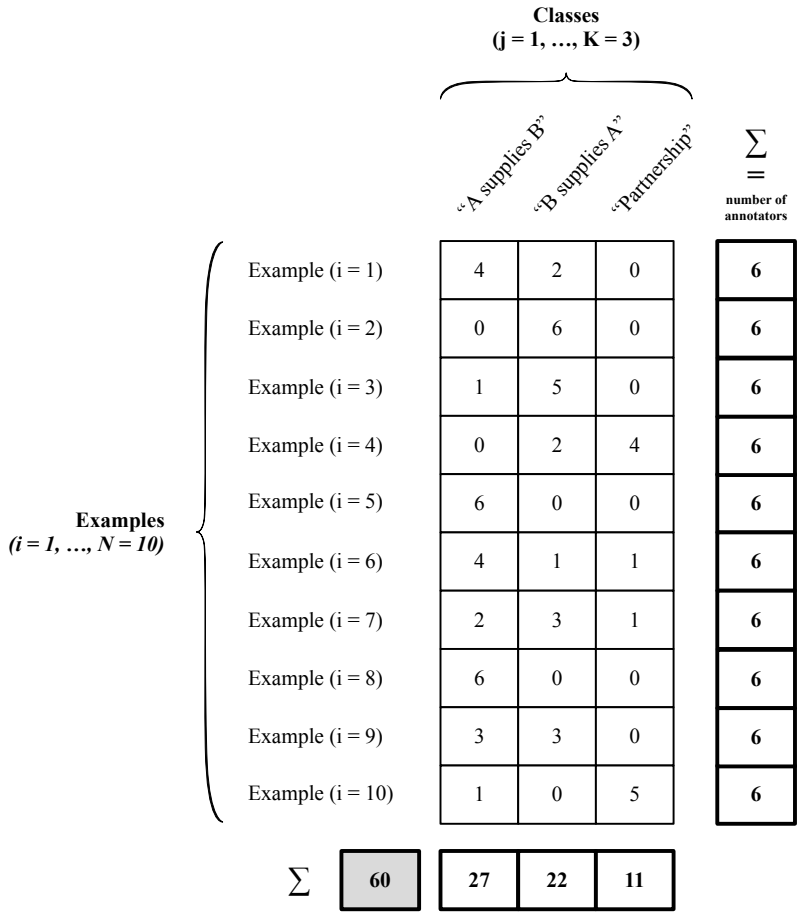

Fig. 4 Matrix for an example calculation of Fleiss' Kappa (fictitious example)

Definition Fleiss' Kappa is defined by the following equation:

$$\text{Fleiss' } \kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{6}$$

The "bar" on top of the variables indicates that the arithmetic mean will be used in the computation of these variables.

Expected agreement ($\bar{P}_e$) $\bar{P}_e$ in the above equation denotes the agreement *expected by chance alone* and is calculated as:

$$\bar{P}_e = \sum_{j=1}^{K} p_j^2 \tag{7}$$

with $p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$ and $\sum_{j=1}^{K} p_j = 1$ and the following notations:

- N denotes the total number of examples, indexed by $i = 1, \ldots, N$
- K denotes the number of classes that annotators have to choose from, indexed by $j = 1, \ldots, K$
- n denotes the number of annotations per example
- More specifically n_{ij} represents the number of annotators who assigned example i to class j.

Observed actual agreement ($\bar{P}$) $\bar{P}$ in the definition of Fleiss' κ denotes the observed agreement and is defined as:

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{K} n_{ij}^2 - Nn\right) \tag{8}$$

Worked example

Expected agreement ($\bar{P}_e$) p_j is the proportion of all examples that were assigned to a specific class j. This corresponds to the ratio of column total to overall number of assigned examples (number of annotations per example times number of examples).

N denotes the total number of examples, in this case $N = 10$. n denotes the number of annotations per example, as is given as $n = 6$. For the case $p_{j=1}$, the result is:

$$p_{j=1} = \frac{1}{Nn}\sum_{i=1}^{N} n_{i1} = \frac{1}{10 \cdot 6}\sum_{i=1}^{10} n_{i1} = \frac{27}{60} = 0.450$$

$$p_{j=2} = \frac{1}{Nn}\sum_{i=1}^{N} n_{i1} = \frac{1}{10 \cdot 6}\sum_{i=1}^{10} n_{i1} = \frac{22}{60} \approx 0.367$$

$$p_{j=3} = \frac{1}{Nn}\sum_{i=1}^{N} n_{i1} = \frac{1}{10 \cdot 6}\sum_{i=1}^{10} n_{i1} = \frac{11}{60} \approx 0.183$$

The sum over all p_j should be equal to 1.

$$\bar{P}_e = \sum_{j=1}^{K} p_j^2 = (\ 0.450\ )^2 + (\ 0.367\ )^2 + (\ 0.183\ )^2 \approx 0.371$$

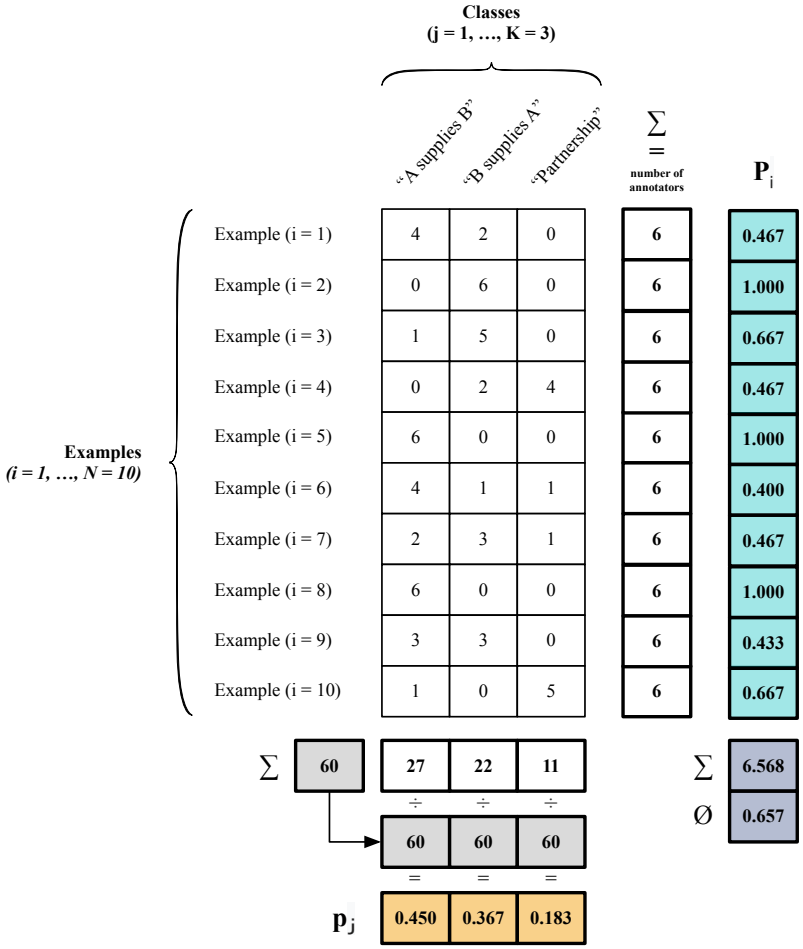

Fig. 5 Matrix for an example calculation of Fleiss' Kappa

Observed actual agreement ($\bar{P}$) To calculate the observed actual agreement, the extent P_i to which annotators agree for the i-th example needs to be calculated first. This corresponds to the ratio of the number of annotator–annotator pairs in agreement and the number of of all possible annotator–annotator pairs.

$$
\begin{aligned}
P_i = & \\
& \frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij}-1) = \\
& \frac{1}{n(n-1)} \sum_{j=1}^{K} (n_{ij}^2 - n_{ij}) = \\
& \frac{1}{n(n-1)} \left[\left(\sum_{j=1}^{K} n_{ij}^2 \right) - (n) \right]
\end{aligned}
\tag{9}
$$

For the first example (first row), P_i is calculated as follows:

$$
\begin{aligned}
P_1 = & \\
& \frac{1}{6(6-1)} \left[\left(\sum_{j=1}^{K} n_{1j}^2 \right) - 6 \right] = \\
& \frac{1}{30} \left[\left(\sum_{j=1}^{K} n_{1j}^2 \right) - 6 \right] = \\
& \frac{1}{30} \left[(4)^2 + (2)^2 + (0)^2 - 6 \right] = \\
& \frac{1}{30} \left[(4)^2 + (2)^2 + (0)^2 - 6 \right] = \frac{1}{30} [20-6] = \\
& \frac{14}{30} \approx 0.467
\end{aligned}
\tag{10}
$$

For the second example (second row), P_i is calculated as follows:

$$
\begin{aligned}
P_2 = & \\
& \frac{1}{30} \left[(0)^2 + (6)^2 + (0)^2 - 6 \right] = \\
& \frac{1}{30} [36-6] = \\
& \frac{30}{30} = 1.0
\end{aligned}
\tag{11}
$$

After all P_i have been computed, $\bar{P}$ can be determined as the mean over all P_i:

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{10}(0.467 + 1.000 + 0.667 + 0.467 + 1.000 + 0.400 + 0.467 + 1.000 + 0.433 + 0.667) = \frac{6.568}{10} \approx 0.657 \tag{12}$$

Overall Kappa

$$\text{Fleiss' } \kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} = \frac{0.657 - 0.371}{1 - 0.371} \approx 0.455 \tag{13}$$

Intra-annotator agreement

Intra-annotator agreement measures the extent to which an annotator labels the same input consistently. This can be considered a proxy to measure both the labelling quality of the annotator as well as the clarity or simplicity of the task itself. Conflicting labels for the same input would indicate labels of poor quality. The same metrics can be used as for inter-annotator agreement. Instead of actually having multiple annotators, the class labels for the same sentence provided by the same annotator at *different points in time* are considered.

Evaluation options and limitations

From the assessment of the labelling quality to the proof of value of an overall approach to automate supply chain mapping from text, there are different abstraction layers that can be evaluated. The longer the NLP processing pipelines get, the larger is the number of potential error sources.

Figure 6 provides a conceptual overview of possible layers and evaluation options one may want to distinguish. The overview is to be read bottom-up; each further layer is supposed to increase the proximity to an actual use case or proof of value. The closer one gets towards the proof of value (higher up the layers in the mental model), the more one is dependent on aspects of information quality and availability that cannot be influenced by the researcher.

The first five evaluation options from the bottom are considered within the scope of the research, while the others are considered out of scope. Each layer shall briefly be discussed to raise awareness about the differences between them. By following the methodology and

evaluating these aspects one would be validating the methodology. The validation options could also be extended from just direct suppliers to sub-tier suppliers.

Labelling feasibility & quality This layer corresponds to the first stage (Section **?? ??**). By measuring inter- and intra-annotator agreement one can answer questions about the labelling feasibility and quality. In this stage, NER errors, information availability and veracity of collected statements are ignored. Error sources are the ambiguity of the language, the task complexity and the instruction quality as well as human errors.

Classification performance This layer corresponds to the second stage (Section **?? ??**). The objective is now to measure how well a trained classifier can distinguish the classes. Because the classifier has been trained on the human-annotated corpus, all errors made during the annotation process will also be reflected in the corpus and the classification performance. In addition to these error sources, a classification error needs to be considered which may be due to the labelling quality as well as the size and bias of the dataset.

Generalisability This layer shall correspond to the question how well the trained classifier works on different datasets. If the corpus is not fully randomly sampled, the classification performance may be high but the classifier may still perform poorly on other datasets due to overfitting. Additional error sources of this layer are errors in NER or sentence segmentation.

Data sparsity This layer corresponds to how much information about supply chain structures can be extracted from a given dataset. This could be, for instance, the number of extracted buyer-supplier relations per unit of text quantity.

Information availability This layer shall correspond to the question how much of a real supply chain can be extracted using the approach. This could be approximated, for instance, as the percentage of a company's 1st-tier suppliers that could be extracted as validated by that company. An additional source of error is now limited information availability: What was not contained in the processed documents (e.g. because it was not reported or just not part of the processed documents) cannot be extracted.

The following aspects are considered out of scope and will not be evaluated:

Veracity This evaluation option refers to the truthfulness of statements. Information may be factually incorrect, outdated or unsubstantiated rumours.

Novelty A further evaluation option is to measure the extent to which the extracted information is new (e.g. to a specific company).

Value in historical use case The evaluation option closest to an actual proof of value is to apply the approach to an actual historical use case using only information that would have been available prior to that event.

Secondary benefits The last layer has to be seen in separation of the other layers. It aims to capture secondary benefits of the proposed approach that could be evaluated, such as being faster, more up-to-date, more cost-efficient etc. than other manual approaches.

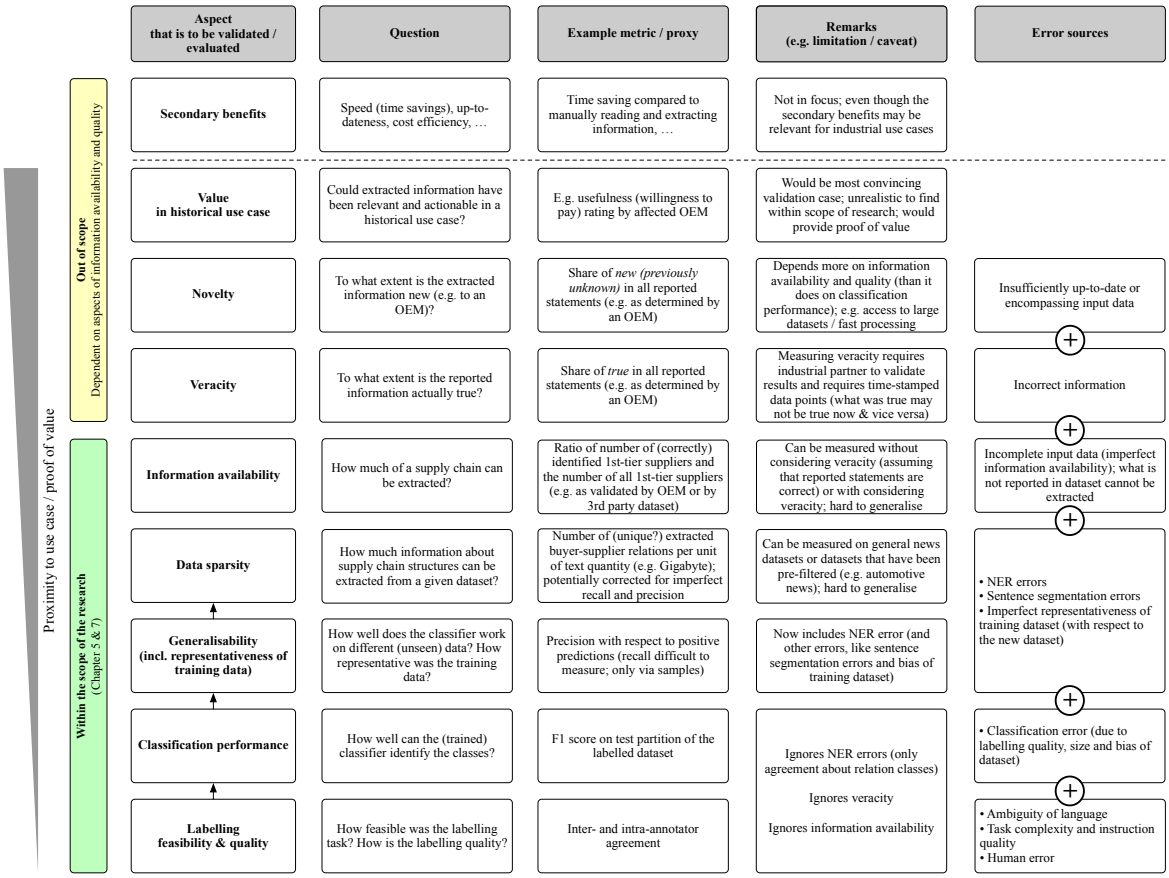

Fig. 6 Evaluation options

References

Elizabeth Anderson. *Extended Value Stream Mapping: Creating a Supply Chain View of Phytosanitary Compliance for Export Timber*. Master's thesis, Lincoln University, 2017. URL https://researcharchive.lincoln.ac.nz/handle/10182/8714.

Kanda Boonsthonsatit and Siripong Jungthawan. Lean supply chain management-based value stream mapping in a case of Thailand automotive industry. *2015 4th IEEE International Conference on Advanced Logistics and Transport, IEEE ICALT 2015*, (July):65–69, 2015. doi: 10.1109/ICAdLT.2015.7136593.

Martin Christopher. *Logistics and Supply Chain Management: Creating Value-Adding Networks*. Pearson Education Limited, Harlow, 3rd editio edition, 2005.

Martin Christopher and Helen Peck. Building the resilient supply chain. *International Journal of Logistics Management*, 15(2):1–13, 2004. ISSN 0957-4093. doi: 10.1108/09574090410700275. URL http://dx.doi.org/10.1108/09574090410700275.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Joseph Fleiss. Measuring Nominal Scale Agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Kevin A Hallgren. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23–34, 2012. ISSN 1913-4126. URL http://www.ncbi.nlm.nih.gov/pubmed/22833776{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3402032.

Uta Jüttner and S Maklan. Supply chain resilience in the global financial crisis: An empirical study. *Supply Chain Management*, 16(4):246–259, 2011. ISSN 1359-8546. doi: 10.1108/13598541111139062.

Yusoon Kim, Yi Su Chen, and Kevin Linderman. Supply network disruption and resilience: A network structural perspective. *Journal of Operations Management*, 33-34:43–59, 2015. ISSN 02726963. doi: 10.1016/j.jom.2014.10.006. URL http://dx.doi.org/10.1016/j.jom.2014.10.006.

Klaus Krippendorff. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112, 2011.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

Timothy J Pettit. *Supply Chain Resilience: Development of a Conceptual Framework, an Assessment Tool and an Implementation Process*. Doctor of philosophy, The Ohio State University, 2008.

Serhiy Y. Ponomarov and Mary C. Holcomb. Understanding the concept of supply chain resilience. *The International Journal of Logistics Management*, 20(1):124–143, 2009. ISSN 0957-4093. doi: 10.1108/09574090910954873.

Michel E Porter. *Competitive advantage: Creating and Sustaining Superior Performance*. The Free Press, New York, first free edition, 1985.

James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Sebastopol, CA, 1st edition, 2013. ISBN 9781449306663.

Mike Rother and John Shook. *Learning to see: Value stream mapping to create value and eliminate muda*. The Lean Enterprise Institute, Brookline (US), 1.2 edition, 1999.

William A Scott. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325, 1955.

Yossi Sheffi. *The Resilient Enterprise: Overcoming Vulnerability for Competitive Advantage*, volume 1. The MIT Press, 2005. ISBN 0262195372.

Manuel F. Suarez-Barraza, José Miguel-Davila, and C. Fabiola Vasquez-García. Supply chain value stream mapping: a new tool of operation management. *International Journal of Quality and Reliability Management*, 33(4):518–534, 2016. ISSN 0265671X. doi: 10.1108/IJQRM-11-2014-0171.