Vol.2016-IFAT-122 No.5
Vol.2016-DC-101 No.5
2016/3/24

翻訳プロダクトの品質評価に向けた翻訳業界の取り組み

西野 竜太郎†

概要： 翻訳の品質評価をする際，作業の「プロセス」に注目する場合と，翻訳成果物である「プロダクト」に注目する場合とがある．後者のプロダクト評価では，従来，産業界ではエラーベースの手法が主流だったが，学術界からはいくつかの面から批判もあった．しかし批判に応えるような形で，2010 年代以降，業界団体から MQM や DQF といった評価方法が提案され始めた．それらの方法は共通化が進みつつあり，課題は残りつつも国際的な基準に発展する可能性も考えられる．

Efforts by Translation Industry to Build Quality Evaluation Methods for Translation Products

RYUTARO NISHINO†

1. はじめに

どのような翻訳が「良い」かについてはさまざまな意見があるだろう．実際，翻訳物（プロダクト）の品質とその評価方法については，産業界でも学術界でも論争が絶えず[1][2], 広く受け入れられる共通基準は存在しないと考えられる．翻訳作業の「プロセス」については 2015 年に ISO 17100 が発行され，ある程度の国際的な合意が存在するものの,「プロダクト」の品質評価については関係者による努力が待たれるという状況である．本論では，プロダクト評価の問題点，産業界と学術界における方法，および産業界における取り組みについて紹介する．

2. 翻訳業界の現状

2.1 関係者と評価

共通基準が存在しないとはいえ，現在でもプロダクトの評価自体は日常的に行われている．評価をしなければ，例えば翻訳者からの納品物を翻訳会社が検収することはできない．現在の翻訳業界における典型的なプロダクトの流れとそれに伴う評価の方向を図 1 に示す[a]．関係者はソース・クライアント，翻訳会社，翻訳者，最終読者という 4 者が基本で, 最初の 3 者は翻訳の「生産側」, 最終読者は「消費側」と区分することもできる．図で実線はプロダクトの流れ，点線が評価の方向を指す．ソフトウェア・マニュアルの和訳を例にしてみよう．まずソフトウェア会社（ソース・クライアント）は翻訳会社に和訳を発注する．それを受けた翻訳会社はさらに翻訳者（多くが在宅フリーランス）に仕事を依頼する．翻訳者は翻訳作業を終える[b]と，翻訳会社にプロダクトを納品する．このとき翻訳会社は翻訳者からの納品物を評価する（図中の A）．さらにそれがソフトウェア会社に戻され, ここでも評価が発生する（図中の B）．ソフトウェア会社はマニュアルをウェブ上などに掲載してユーザー（最終読者）に読んでもらうが，ここでもアンケートなどの形で評価が発生することがある（図中の C）．つまり，あるプロダクトに対し，別々の主体が計 3 回の評価をしていることになる．

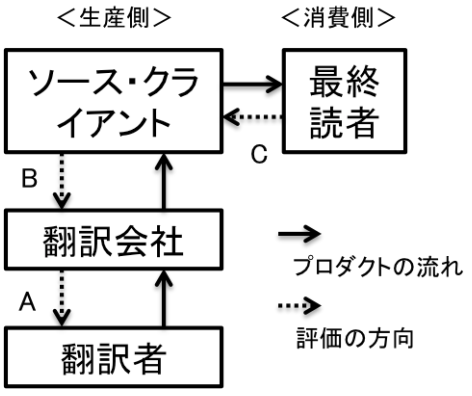


図 1 翻訳プロダクトと評価の流れ

では評価の目的は何だろうか？ プロダクトを受け取った人が自分にとって満足できるものかどうかを判定するために実施することが多いだろう[c]．この判定基準が明示されていれば納品物に関するトラブルが減る．さらに統一された共通基準であれば，二者間で細かな部分までいちいち合意を取るコストもかからず，ビジネスは円滑化される．

† フリーランス翻訳者（ryutaro@nishinos.com）
Freelance translator

a これは実務翻訳や産業翻訳と呼ばれる分野（IT，工業，医薬，特許など）の例であり，書籍や映画字幕などの分野では事情が異なることもある．

b 翻訳者が作業中に自己評価をすることがあるが，今回は関係者間における評価方法の違いをテーマにするため，ここでは省略する．

c 他にも，翻訳者の能力測定や，訳文改善を意図したフィードバックを目的に評価することもあるだろう．

Vol.2016-IFAT-122 No.5
Vol.2016-DC-101 No.5
2016/3/24

翻訳業界でプロダクトの品質基準が望まれる最大の理由はこれらの点であると思われる．

2.2 プロダクト評価における問題点

プロダクト評価における大きな問題点として 2 つが挙げられる．まずそもそも「品質」が何を指すのか立場によって異なり共通了解がない点，次に定評のある評価方法がない点である．

2.2.1 立場で異なる「品質」の内容

翻訳業界には基本的にソース・クライアント，翻訳会社，翻訳者，最終読者という 4 者がいると述べた．あるソフトウェア・マニュアルの和訳について，それぞれ例えば以下のように品質評価をしたとする．

- 最終読者：「マニュアルで『である調』は読みにくい．だから低品質」
- ソース・クライアント：「社内スタイルガイドに従って『である調』にした．だから品質基準は満たしている」
- 翻訳会社：「この納期と料金で対応できるのはここまで．だから質は十分」
- 翻訳者：「この分野で 10 年の翻訳経験がある．だから品質には自信がある」

どの意見にも一理ありそうではあるが，読みやすさ，社内基準，納期と料金，翻訳経験とさまざまな要素が登場していて，品質が何を指しているのか共通しているようには思えない．この 4 者が品質について話し合っても噛み合わないだろう．筆者の経験上，こうした不一致は見られる．翻訳は言葉が関係するため，詩や小説などとは程度の違いはあるだろうが，実務翻訳であっても美醜などの主観的判断は排除できない（ここに工業製品との差があると思われる）．さらに複数の独立した事業主体（ソース・クライアント，翻訳会社，翻訳者）が関わるため経済的利害が絡むこともある．つまり，現在は立場によって「品質」の指す内容が異なり，生産的な議論が容易ではない状態にある．

2.2.2 定評のある評価方法がない

プロダクトの評価方法は，産業界にも学術界にも複数存在する．ただしどの方法に対しても批判や不満があり，定評のある評価方法はない．業界で統一された共通基準が未だに存在しないのも当然だと言える．

次のセクションから産業界および学術界における評価方法をいくつか紹介する[d].

d 機械翻訳の「自動評価」は扱わない．自動評価は基本的に人間の作った参照訳にどれだけ近いかで評価をするが，本論ではその参照訳自体の質を話題としている．また自動評価が前提としているであろう「等価」も前提としていない．

3. 産業界における評価方法

現在ローカリゼーションを中心とした翻訳業界では「エラーベース」の方法が用いられることが多い[1][3]．2011 年に解散した LISA（Localization Industry Standards Association）という業界団体が 1980 年代に開発した手法（LISA QA モデル）が有名である．各企業独自の評価モデルの多くはこれから派生しているとされる[3]．エラーベースの方法では，評価者は翻訳に含まれるエラーをカテゴリー（誤訳，用語，スタイルなど）別に見つける．このとき，各エラーで重大度に応じて点数（例 : Critical 10 点，Major 5 点，Minor 1 点）を付ける．翻訳全体に含まれる合計点数がある値を超えると不合格にするというものである．不合格を繰り返すと，例えばソース・クライアントは当該翻訳会社への発注を取りやめるといった決定を下す．何をエラーとするか，エラーの点数は何点とするかはカスタマイズ可能である．同様のエラーベースの方法は自動車産業の翻訳でも用いられている（SAE J2450）．

3.1 エラーベースに対する批判

このエラーベースに対する批判がいくつかある．

まず，経験的で理論化されておらず，ある組織（企業や団体など）内での使用には十分かもしれないが，一般化できないという批判である[4]．確かに同じエラーベースではあっても，エラー分類や重大度の点数付けが組織ごとに異なることはあるだろう．

次の批判として，語や文のミクロレベルを中心に分析的に見ており，文章（テクストとも）全体のマクロレベルに対する意識が薄いという批判である[1][2]．エラーベースで見つかるのは例えば用語違反やスタイル違反であり，文章全体（例えばマニュアル）がそのジャンルにおける慣習に従って翻訳されているかどうかは見落としがちということである．

さらに，単一の基準（「one-size-fits-all」とも）ではさまざまなコンテンツ種類（例：特許文書，マーケティング資料）や状況などに応じて柔軟に対応できないという批判もある[3]．原文への忠実さが求められる文書と読者に訴求する文書とでは，評価方法に違いがあってもおかしくはない．テクスト内のエラー数だけであらゆる翻訳を評価するのは無理があるという意見は理解できる．

4. 学術界における評価方法

学術界ではさまざまな評価方法が提唱されている．ここでは Colina の分類[2]に基づき，等価ベースと非等価ベースに分けて説明する．

Vol.2016-IFAT-122 No.5
Vol.2016-DC-101 No.5
2016/3/24

4.1 等価ベースのアプローチ

等価（equivalence）とは，原文と訳文[e]との間に「等しい価値」があることを指す[5]．翻訳学においては論争の多い用語であるが，等価ベースのアプローチはこの等価に注目して評価する．

まず読者反応アプローチでは，読者が原文と訳文に対して同じような反応を示すかどうかで品質を判断する[2]．このアプローチには，ある種の質（情報の明瞭性など）は測れたとしてもそれを全体の質にすることはできないなどの批判[6]はあるが，「読者」を認識した点は評価されるべきともされている[2]．

続いてテクスト・アプローチで，文章全体（テクスト）の種類や機能といった点に注目する．例えば House の方法では，原文テクストと訳文テクストの言語的状況的特徴を分析（レジスター分析）し，その両者が機能的に等価であるかを評価する[6]．語や文のミクロレベルではなく，文章全体のマクロレベルで評価をしている点が特徴である．

4.2 非等価ベースのアプローチ

非等価ベースのアプローチでは原文を必ずしも絶対視せず,「訳文としてどうか」という視点から評価する傾向にある．

最も有名なのは「機能主義」である．機能とは目的（スコポス）のことで,「翻訳が何のために使われるのかに焦点を当てる」[7]方法である．そのため，ある訳文がその翻訳の目的に適合しているかどうかが評価基準となる．例えば製品広告の和訳であれば,原文と等価であるというよりも，日本の消費者に訴求できる文章であるかどうかが重要となる．そこで機能主義では翻訳の目的を明示した翻訳依頼者からの指示書（translation brief）が重視される．

さらに機能主義を発展させた Colina の機能主義的コンポーネント・アプローチ[4]がある．これは評価にいくつかのコンポーネント（訳文，機能とテクストの妥当性，非専門的内容，専門的内容と用語）を用意し，翻訳依頼者のコミュニケーション目的に応じて優先順位を付けて使用する方法である．語や文レベルではなく，文章全体で評価する．

4.3 学術界のアプローチに対する批判

学術界で提唱された評価方法は，現実には実務で使われていない[8]．これは，学術界が知的好奇心を満たそうとするのに対して産業界では実践的な方法を望むためという理由[8]や, 学術界が品質の特定の側面（テクストの等価など）しか見ていない点や実務や教育の場面で適用が難しいことが多い点が理由[4]であると考えられている．実務では納期や予算といった制約も考えなければならないし，評価作業に手間がかかり過ぎても利用されないだろう．ただし，翻訳の目的などを考慮に入れた機能主義が登場し，徐々に産業界に近づいているともされる[8]．

5. 産業界での新たな取り組み

2.2.1 でプロダクト評価における問題点として（1）立場で「品質」の内容が異なる，(2）定評のある評価方法がない，という 2 点を挙げた．こういった問題に対する産業界での新たな取り組みを紹介する．

5.1 「品質」の分類と明確化

2014 年に翻訳関係の研究者と実務家の数人は，アメリカの経営学者 Garvin の議論を参考に，翻訳の品質に関する 5 つのアプローチを提示した[9]．この 5 つを順に説明する．

まず「超越的」アプローチである．もともと哲学の言葉だが，良い文章に多く触れるといった経験を通して培われた力で，品質の良し悪しを直観的に判断することを指す．

次に「プロダクトベース」で，製品やサービスの品質は原材料や特質で測定可能というアプローチである．例えば Garvin は品質の高いアイスクリームには乳脂肪分が多く含まれるという例を挙げている[10]．ただしこのアプローチで品質を上げようとするなら当然費用も高くなるだろう．

3 つめは「ユーザーベース」である．製品やサービスの品質は，ユーザーのニーズ，要望，好みを満たしている度合いによって決まるという考え方で，消費側からの見方だと言えよう．

4 つめの「生産ベース」[f]は，あらかじめ定めた要件や仕様をどの程度満たしているかで品質が決まるというアプローチである．ユーザーベースとは逆に，生産側からの見方となる．

最後は「価値ベース」で，費用と便益（cost and benefit）で品質を測定するアプローチである．便益が費用に比べてより大きいならば製品やサービスにより価値があり，そのため品質もより高いと見なす．プロダクトベースの測定で品質が高いと判断がなされても，もし費用も大きければ，この価値ベースでは品質は相対的に低くなる．

2.2.1 で立場の違いによる品質評価の違いの例を挙げた．これは上記 5 つのアプローチのいずれかに当てはまりそうである．以下に立場の違いと，そのアプローチを矢印の後に書く．

- 最終読者：「マニュアルで『である調』は読みにくい．だから低品質」 → ユーザーベース
- ソース・クライアント：「社内スタイルガイドに従って『である調』にした．だから品質基準は満たし

e 原文は ST（source text や start text），訳文は TT（target text）と呼ばれることが多いが，ここでは「原文」と「訳文」としておく．

f 英語では production-based である．Garvin のオリジナルでは manufacturing-based だが, 翻訳場面に合わせて production-based とされた[9].

ている」　→　生産ベース
- 翻訳会社：「この納期と料金で対応できるのはここまで．だから質は十分」　→　価値ベース
- 翻訳者：「この分野で 10 年の翻訳経験がある．だから品質には自信がある」　→　超越的

確かに 5 つのアプローチを使うと，立場の違いによる品質評価の違いがうまく分類できるのではないかと考える．ただしまだ「立場が違うと評価方法も違うが，その違いは 5 つほどに分類できそうだ」と分かった段階である．具体的な翻訳品質についてさらに議論を深める必要があるが，少なくとも議論の共通基盤が提出された点は大きな一歩として評価できるだろう．

5.2 批判を取り込んだ評価方法

3.1 で，産業界で主に用いられているエラーベースの方法に対し，学術界から批判があると述べた．一般化できない点，マクロレベルへの意識が薄い点，さらにコンテンツ種類などに応じた柔軟な対応ができない点であった．2010 年代以降に産業界で提唱された評価方法は，こういった批判も取り込んで改善を図ろうとしている．以下で 2 つを紹介する．

5.3 MQM

MQM（Multidimensional Quality Metrics）は，翻訳テクストの評価および翻訳テクストにおける問題の識別に使用する品質メトリクスを記述および定義するフレームワークである[11]．ドイツ人工知能研究センター（DFKI）および QTLaunchPad が開発し，2014 年 2 月に最初のバージョン 0.1，2015 年 12 月に 1.0 が公開されている．

MQM はすぐにそのまま評価作業に適用できるものではない．まずさまざまな要件に応じて自分の品質評価メトリクスを作成し，その後で評価作業に適用するという流れになる．

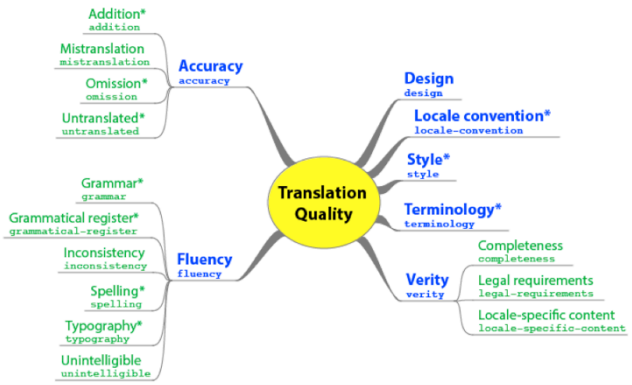
メトリクス作成でまず考慮するのは評価方法で，分析評価（Analytic）と全体評価（Holistic）が中心的である．分析評価では従来の LISA QA モデルと同様，ミクロレベルでエラー数をカウントしてスコアを出す．そのため時間と労力がかかる．一方，全体評価では文章全体に対して評価を実施する．例えば最終読者に全体的な印象を尋ねるといった場面で使える．マクロレベルを含めたという点では学術界の批判に応えていると言えよう．

続いて考慮するのは品質側面（Quality dimensions）である．具体的には以下のイシュー・タイプ（エラーの種別）から選択する．

- Accuracy（原文と訳文の関係．誤訳など）
- Fluency（訳文の形．文法ミスなど）
- Design（テクストの物理的な表示）
- Locale convention（桁区切りなどの地域慣習）
- Style（スタイルガイドなど）
- Terminology（分野や組織の用語）
- Verity（最終読者などに適した内容）

これらのイシュー・タイプは細分化されており，どの粒度で評価するかも選べる．図 2 に MQM Core と呼ばれる中核的な項目を示しているが，シンプルに MQM を適用したい場合はこの Core だけメトリクスに含めて利用できる[g]．例えば左上に「Accuracy」とある．この Accuracy のレベルで評価することも，Accuracy をさらに細分化した「Addition」（追加）や「Mistranslation」（誤訳）というレベルで評価することもできる．

どの側面をどの粒度で評価するか選択できるため，ある程度の柔軟性を持っている．例えばソース・クライアントが翻訳会社からの納品物を検収する場合は，分析評価を用いて全側面を全レベルで確認できる．一方，最終読者にアンケートを取ってマニュアルを評価してもらう場合は，全体評価を用いて Fluency のみを確認してもらうことも可能である．



（http://www.qt21.eu/mqm-definition/definition-2015-12-30.html より引用）

図 2　MQM Core

簡単にまとめると，MQM はイシュー・タイプ（エラー項目）を柔軟に構成し，テクストを分析評価または全体評価する方法と言える．

5.4 DQF

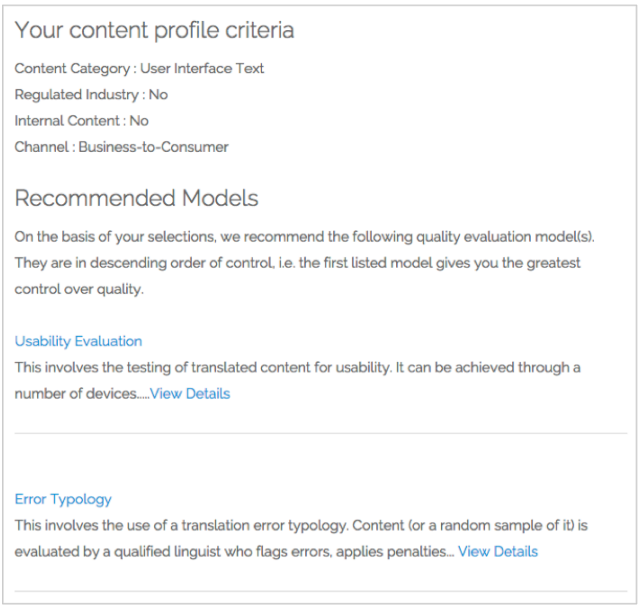
DQF（Dynamic Quality Framework）は TAUS（Translation Automation User Society）という翻訳業界団体が開発した品質評価フレームワークである．2012 年から公開された[12]．

特徴的なのは，実際の評価前にまず「コンテンツ・プロファイリング」を実施する点である．この結果，翻訳対象の種類（マニュアル，マーケティング資料，ソーシャル・

g 状況に応じてさらに Core から選択することも可能．ただし「Accuracy」と「Fluency」の 2 つは最低限含めることが推奨されている[11]．

メディアなど）やコミュニケーション・チャンネル（社内向け，企業-消費者，など）に基づいて評価方法がいくつか推薦される．具体的には，従来からのエラーベースの評価に加え，ユーザビリティー評価，忠実さ／流暢さ，コミュニティー評価，リーダビリティー評価などが推薦される．評価者は推薦項目から重要と思われるものを選択し，実際の評価作業に入る．従来のようにエラー一辺倒ではなく，コンテンツ種類などに応じて柔軟に評価できるため「動的」（dynamic）[3]であるとされる．またユーザビリティーのようにテクスト外の項目も入れている点が特徴的である．

コンテンツ・プロファイリングのウェブサイト[h]で，例えばコンテンツは「User Interface Text」，規制業界かどうかは「No」，社内向けかどうかは「No」，チャンネルは「Business-to-Consumer」で試すと，図 3 のようにユーザビリティー評価（Usability Evaluation），エラー種別（Error Typology）といった評価項目が推薦される．項目は品質に影響を与えやすいと考えられる順に提示される．



（https://evaluate.taus.net/evaluate/content-profiling/profile-your-content より引用）

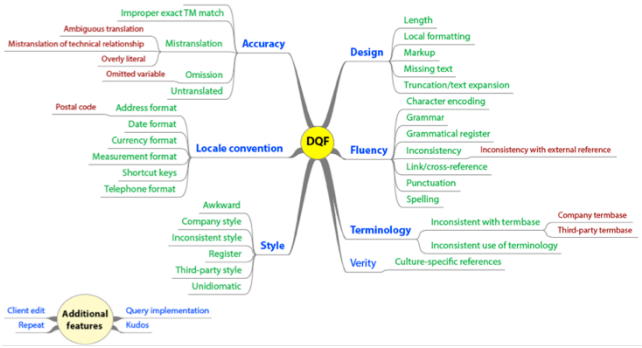
図 3　DQF のコンテンツ・プロファイリング結果例

簡単にまとめると，DQF はコンテンツ種類に応じて動的に複数のアプローチから評価する方法である．従来からあるテクスト内のエラー数で評価するアプローチも，テクスト外のユーザビリティーなどで評価するアプローチも含まれる．

5.5 進む共通化と課題

このように 2010 年代以降，批判に応えるような形で複数の団体が新しい評価方法を提案している．マクロレベルを意識し，状況に応じて柔軟に評価できる方法である．

こうした団体間での協力も進みつつあり，2015 年に MQM のイシュー・タイプと DQF のエラー種別は共通化された[13]. 図 4 に共通化後の DQF のエラー種別を示すが基本的に DQF が MQM に用語を合わせている．他の団体との協力がさらに進めば，より一般性の高い国際的な基準になる可能性も考えられる．



（http://www.qt21.eu/mqm-definition/definition-2015-12-30.html より引用）

図 4　MQM と共通化された DQF のエラー種別

ただし課題もある．比較的新しい手法のため，実運用の実績やノウハウが乏しい点である．実運用が難しければ学術界からの提案に対するのと同じ批判が発生し得るし，利用者も増えないだろう．また，さまざまな立場の人に受け入れてもらう努力もより一層求められるはずだ．

6. 結論

翻訳プロダクトの評価に関する翻訳業界の取り組みについて紹介した．従来，産業界ではエラーベースの手法が主流だったが，それに対しては学術界からの批判もあった．そのような批判を取り込むような形で，2010 年代から MQM や DQF といった評価方法が提案され始めた．こういった方法は共通化が進みつつあり，課題もあるが国際的な基準に発展する可能性も考えられる．

参考文献

[1] Jiménez-Crespo, M.Á. Translation and Web Localization. Routledge (2013).
[2] Colina, S. Evaluation/Assessment. In Y. Gambier and L. van Doorslaer, eds., Handbook of Translation Studies. John Benjamins Publishing Company, pp.43–48 (2011).
[3] O'Brien, S. Towards a dynamic quality evaluation model for translation. The Journal of Specialised Translation, 17 (2012).
[4] Colina, S. Fundamentals of Translation. Cambridge University Press (2015).
[5] Pym, A. Exploring Translation Theories (Second Edition). Routledge (2014).
[6] House, J. Quality. 翻訳研究のキーワード. 研究社 (2013).
[7] 武田珂代子. 機能主義的アプローチ（スコポス理論）. よくわかる翻訳通訳学. ミネルヴァ書房 (2013).

h こちらの URL から利用可能（2016 年 2 月時点）：
https://evaluate.taus.net/evaluate/content-profiling/profile-your-content

[8] Drugan, J. Quality In Professional Translation: Assessment and Improvement. Bloombury Publishing (2013).
[9] Fields, P., Hague, D., Koby, G.S., and Melby, A. What Is Quality? A Management Discipline and the Translation Industry Get Acquainted. Revista Tradumàtica, 12, pp.404-412 (2014).
[10] Garvin, D.A. What Does “Product Quality” Really Mean? Sloan Management Review 26, pp.25-43 (1984).
[11] Lommel, A., Burchardt, A., and Uszkoreit, H., eds. Multidimensional Quality Metrics (MQM) Definition v1.0 (2015).
[12] TAUS. TAUS History. https://www.taus.net/history (2015).
[13] TAUS. DQF and MQM Harmonized to Create an Industry-Wide Quality Standard. https://www.taus.net/think-tank/news/press-release/dqf-and-mqm-harmonized-to-create-an-industry-wide-quality-standard (2015).