

『グリム童話』と『ドイツ伝説集』に含まれる 物語の品詞分解による再分類

かた やま こうじろう
片 山 耕二郎

本論文は、グリム兄弟が編纂した『子どもと家庭のためのメルヒェン集』（初版 1812、第 7 版 1857、以下『グリム童話』と記す）と『ドイツ伝説集』（1816-1818）所収の物語について、単語数・語彙数と品詞数を機械的に分析し、その結果に基づいた機械学習によって再度、メルヒェンと伝説に分類することで、データサイエンスを用いた文学研究の可能性を検討したものである。

ドイツ語の品詞分解には Christian Wartena の開発した、Python 用ライブラリ The Hanover Tagger (HanTa) の ver. 1.1.1 を用いる。単語のトークン化（文から単語への切り分け）には言語処理ライブラリ NLTK の ver. 3.7 を用いる。これらの詳しい使用手順は、拙論「The Hanover Tagger による品詞分解を用いた「ゲーテ時代」文学研究序論」で述べた⁽¹⁾。また、機械学習にはライブラリ scikit-learn の ver. 1.2.2 を用いる。

1. グリム兄弟によるメルヒェン・伝説収集と分類基準について

グリム兄弟は 19 世紀前半に活躍したドイツのロマン主義者であり、言語学・法学等に通暁したゲルマニストとして高名である。法学者サヴィニーの元で法学を学んだのち、ドイツ法のみならずドイツ文化への関心を強く抱き、またナポレオン支配地域で図書館員として働いた経験から、ドイツ文化の保護とドイツの統一を願った。さらに、同様の意図を持つアルニムやブレンターノの働きかけを受けた。主としてこうした動機から、彼らの業績として最も有名な『グリム童話』のほか、『ドイツ伝説集』や『ドイツ語辞典』も生まれている。したがって、彼らにとってメルヒェンと伝説はまず、ドイツ文化を受け継いだ口承文芸として共通性を持つものである。

『ドイツ伝説集』の序文でグリム兄弟は両者の共通性を述べている。

メルヒェン、伝説、歴史、[中略] これらは互いに並びあっていて、相前後しながら、太古の時代を爽やかで生き生きとした精神として身近にもたらそうとする。(Grimm, 1816-1818, p. 5)⁽²⁾

両者[メルヒェンと伝説]の類似性が見逃されたり、両者がどこまでも混ぜ合わさり、多かれ少なかれ似たものになっていることが否定されたりしてはならない。メルヒェンと伝説は、感覚的に自然なものや理解可能なものを理解不能なものに常に混ぜ合わせている限りにおいては、歴史[中略]とは対照的である。(Grimm, 1816-1818, p. 7)

歴史、伝説、童話という分類は明らかに認められているもので、無視することはできない。それでも三つのどれに位置づけるか決められないこともある。たとえばホッラ婦人は伝説にも童話にも現れるし、伝説のような状況が、かつて史実として起きたということもありうるのだ。(Grimm, 1816-1818, p. 15-16)

このように、グリム兄弟にとってメルヒェンと伝説は区別可能なジャンルであるが、境界的な作品もある。

ではグリム兄弟は何を根拠に両者を区別したのか。基本的には内容による。ふたたび引用する。

それぞれには独自の領域がある。メルヒェンはより詩的で、伝説はより歴史的である。メルヒェンはほぼ自らのみに立脚し、生まれもった繁栄と完成に至る。伝説は、色彩の多様性に乏しいという特徴のほかに、知られているもの、馴染みあるもの、つまり場所や、史実となった名前に結びついているという特徴がある。(Grimm, 1816-1818, p. 5-6)

メルヒェンは詩的であり、伝説は歴史的である。つまり、前者は物語そのものの面白さがジャンルを規定するのに対し、後者は物語が土地や歴史に結び付けられ、そこで起こった出来事だと聞き手に信じさせることに重きを置く。この区別は、人間が直感的に行いうるものだが、機械によっても分析に用いうる基準についてはどう述べているか。

ほとんど唯一、メルヒェンだけが、(いにしえのヒルデブラントのように自らにおいて有名になり、一般的になった名前を除いて)固有名なしで、ドイツ古来の英雄伝説

の一部を保つことができたのに対し、わが民族の歌謡や伝説には、最古の時代に由来する無味乾燥な個々の人名、地名や習俗がこれほどにも多く、しっかりと残っているのである。(Grimm, 1816-1818, p. 6-7)

固有名詞がほぼないのがメルヒェンで、それらがしっかりと残っているのが伝説である。したがって、機械による分類を試みる場合、固有名詞の割合からメルヒェンと伝説を区別できるのではないかという見通しは立つ。しかし、後述するように HanTa による品詞分解を用いた機械学習による分類基準としては上手くいかない。本論文では、固有名詞に限定せず、品詞全体の割合を活用することで、高い精度での分類を試みる。

2. 機械学習を用いた分類について

データを収集する方法については上述の拙論で述べた。そうしたデータ自体、作品分析のための資料として極めて有用であるが、数値化したデータはまた、機械学習に用いるのにも好適である。ここで簡潔に、機械学習による分類について本論文に役立つ範囲で一般的な説明を行う。

本論文で行うのは、「教師あり学習」による分類である。これは、機械に学習のためのデータを与える際、それがどう分類されるべきかを一緒に教え、そこから分類の仕方を考えてもらうという方法である。しばしば挙げられる例としては、花びらの長さなど（これを説明変数や特徴量と呼ぶ）を元に花の種類（目的変数やラベルと呼ぶ）を判断できるように学習させるものがある。機械はたくさんの花についてその種類と特徴づける数値を教わり、数値をどう用いれば高い確率で種類を判断できるか（例えば花びらの長さが一定以上ならアヤメというように）考える。

本研究が目指すのは、同様の手続きによって、『グリム童話』と『ドイツ伝説集』の物語について、それぞれの総単語数や語彙数、品詞の割合（説明変数）を教えることで、メルヒェンか伝説か（目的変数）を判断させることである。

なお、この判断を学ばせることを訓練と呼ぶが、訓練による分類の精度を調べるために、訓練用データとは別のテスト用データをあらかじめ確保しておき、訓練後にこれを実際に分類させ、正しいラベルと一致するか検証する。この両データはできるだけ質的に等しいことが望ましく、通常はデータをランダムに分割する。ただし、本論文の場合、とりわけ『ドイツ伝説集』は、土地ものと歴史ものなどの区別をもとに類話が並んでいることから、本のどの部分からも均一の割合でデータを確保するのが望ましい。このため、後述するとおり、それぞれの本の物語を登場順に交互に振り分けている。

機械学習による分類では、すでに様々なアルゴリズムが考え出されている。これらについて自分で実装する必要はなく、Python のライブラリ scikit-learn からそれぞれのアルゴリズムのモデルを呼び出し、データを与えるだけで、そのアルゴリズムによる学習をし、分類を行ってくれる。後述するように本論文では5つのアルゴリズムの結果を比較している。

3. 品詞分解結果と人間の目による考察

使用した『グリム童話』『ドイツ伝説集』のデータは、Projekt Gutenberg-DE のものである。『グリム童話』末尾の「子供の聖者伝」については、メルヒェンと伝説の分類実験に相応しくないため除いた。また、テキストの内容について精査は行っていない。大量にデータを処理することで、小さなミスは無視できるものになるのが、こうした研究の一つの長所である。訓練用データとテスト用データを分けるに当たっては、それぞれの作品集をダウンロードしたのち、物語ごとの区切りで1話ずつ振り分けた。基本的には奇数番を訓練用データ、偶数番をテスト用データにしたが、怠け者を扱った『グリム童話』151話には類話が1話収録されているため、ここを境に奇数と偶数が入れ替わっている。メルヒェンがそれぞれ101話と100話、伝説が293話と292話である。以下、訓練用のメルヒェンの集合をM1、テスト用をM2、訓練用の伝説の集合をS1、テスト用をS2と呼ぶ。

M1とS1につき、それぞれの集合全体での品詞割合を主要な品詞についてまとめたの

表 1

	単語数	語彙数	普通名詞	固有名詞	付加形容詞	叙述形容詞	副 詞
M1 総計	152,294	12,606	11.85	3.13	2.58	1.86	6.29
中央	1,219	443	12.23	1.21	2.59	1.82	6.50
平均	1,507.86	467.75	12.18	2.89	2.68	1.89	6.41
S1 総計	95,562	13,218	15.36	3.13	3.47	2.04	5.43
中央	243	151	15.72	2.86	3.49	1.89	5.02
平均	326.15	179.97	15.95	3.53	3.73	1.91	5.23
	前置詞	冠 詞	外来語	並列接続詞	人称代名詞	動詞定形	動詞完了形
	4.52	7.84	6.25	4.46	7.54	8.49	1.22
	4.72	8.50	2.85	4.82	7.97	8.77	1.22
	4.51	8.16	5.52	4.57	7.37	8.37	1.21
	6.69	9.14	1.11	4.47	6.05	7.61	2.02
	6.92	9.3	0.32	4.3	5.26	7.4	1.82
	7	9.81	0.97	4.26	5.35	7.23	2.15

が表1である。品詞の数値はパーセンテージである。

総計はM1とS1それぞれに含まれる全ての物語を一まとめにしたものである。すなわち、物語集としての『グリム童話』と『ドイツ伝説集』の半分についての数値だと考えると分かりやすい。品詞の割合については加重平均とも言える。これに対し、平均値はそれぞれの物語の品詞割合の値を求めた上で、その平均を求めており、また中央値は物語をそれぞれ品詞割合の順に並べた上で、一番中央に来る物語の品詞割合である（物語数が偶数の場合は、中央の2篇を足して2で割る）。

3-1 固有名詞と外来語の平均値と中央値の差について

統計調査でしばしば平均値と中央値を両方求める理由は、平均値は外れ値の存在の影響を受けやすいからである。例えば、機械によって測定されたデータにノイズが混ざったり、人間により入力されたデータに桁の間違いが生じたりして、その極端な値が大きな影響を及ぼすことがある。こうした外れ値に強いのが中央値で、極端な数値を示す項目があっても、その項目は中央の項目を1つずらすに過ぎない。したがって、平均値と中央値、そして場合によっては最頻値を共に用いることに意味が生まれる。

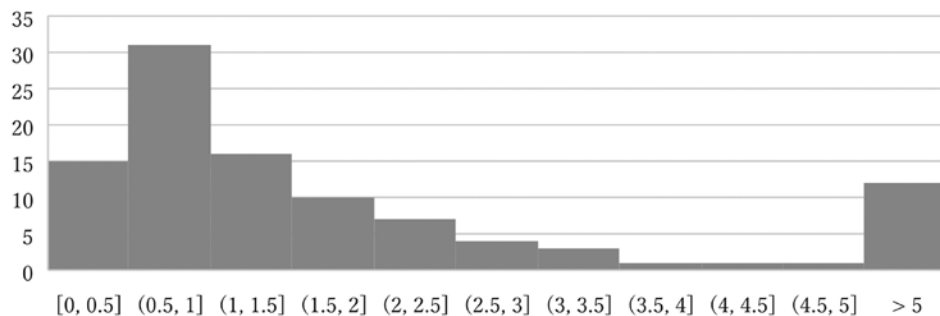
平均値と中央値の違いについて説明したのは、この両者を比較して大きな違いがある場合、しばしば外れ値の影響があり、その外れ値が生まれた理由を考察することでデータの理解が深まるからである。品詞の一覧を見ると、固有名詞と外来語について、メルヒェンの平均値と中央値が大きく異なることが分かる。

M1の固有名詞の割合の中央値は約1.21%で、S1の約2.86%よりかなり低い。グリム兄弟が引用箇所述べていたように、メルヒェンは「固有名詞に触れず」に語られ、伝説には「個々の人名、地名」が含まれていると考えられ、各物語について固有名詞の割合を元にメルヒェンか伝説か区別できそうである。しかし、平均値では約2.89%と約3.53%であるから、そこまで極端な差ではなくなる。なぜメルヒェンの固有名詞の平均値が、中央値より遥かに高くなってしまったのだろうか。

個々の作品の数値を見ると、ときおり固有名詞の割合がきわめて高いものがある。「漁師とおかみ」（約19.8%）、「土の中の小人」（約18.4%）、「もみの木」（約18.4%）、「リンクランク」（約20.0%）などである。これらは方言で語られている物語である⁽³⁾。HanTaは現時点では方言の品詞分解に対応していないため、見覚えのない単語を見つけると、機械的に固有名詞や外来語に分類してしまうのだと考えられる。結果として、固有名詞の割合だけで伝説かメルヒェンかを区別するのは不可能である。

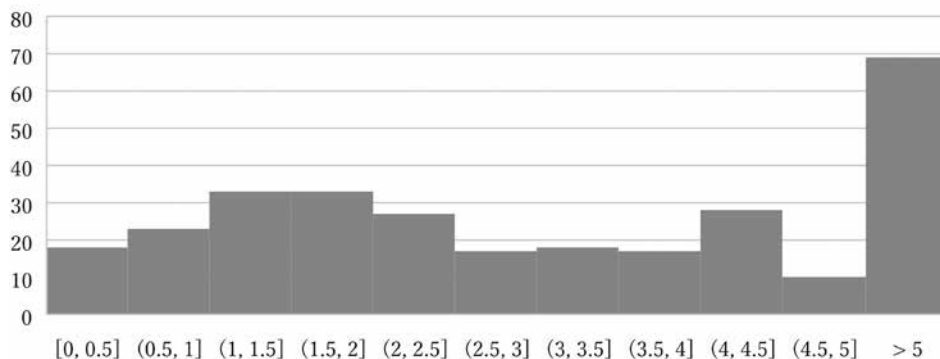
ただし、こうした外れ値がなかったとしても固有名詞だけで判別するのは難しい。次のグラフ1、グラフ2は、M1とS1について、固有名詞の割合を0.5%刻みでヒストグラム

M1の固有名詞の割合のヒストグラム



グラフ 1

S1の固有名詞の割合のヒストグラム



グラフ 2

化したものである。

このグラフからどこかにメルヒェンと伝説の線引きをするなら、1.5%が境になると思われるが、物語数をかぞえるまでもなく、仮に外れ値を除いたとしても、1.5%以上固有名詞のあるメルヒェンも、1.5%未満の固有名詞しか持たない伝説もそれなりに存在する。後述のように HanTa と scikit-learn を用いて複数の品詞をもとに分類した場合、少なくとも 90% 以上の精度で判定できるため、固有名詞に限定するより、他の品詞のデータも材料にした方が優れた判断を示すことが分かる。

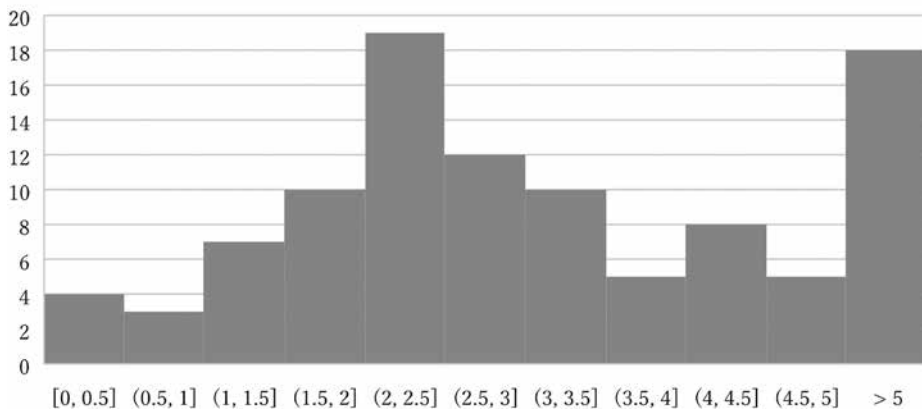
次に外来語について考える。M1 の外来語の割合は中央値が 2.85%、平均値が 5.52% であり、S1 の外来語の割合は 0.32% と 0.97% である。やはり、中央値と平均値に開きがある。固有名詞の場合と同様に、方言で書かれた作品が外れ値として混ざり込んでいるのが主たる理由である。

外来語の値については M1 と S1 の差が大きいために判別の際に有用であると考えられ

る。とりわけ伝説については外来語の割合が0%のものが多い。グリム兄弟はメルヒェンも伝説もドイツ由来のものを集めようと試みたが、メルヒェンは土地に縛られないため、『グリム童話』にはフランス移民などに由来する多くの他地域・言語の物語が混ざり込んでしまった。これに対して伝説は、民族の土地や歴史に結びついているため、ドイツにまつわる伝説はよりドイツの言葉で語られやすい。このため、メルヒェンと伝説では外来語の割合に差が出るものと考えられる。

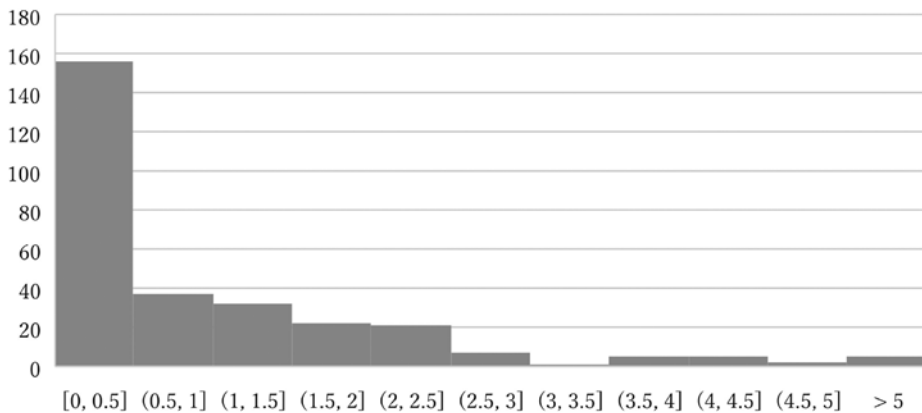
上述の固有名詞のときと同様に、M1とS1の外来語の割合をヒストグラム化したものが次のグラフ3、グラフ4である。

M1の外来語の割合のヒストグラム



グラフ 3

S1の外来語の割合のヒストグラム



グラフ 4

グラフを比較すると、1.5%を境目として判定すれば、メルヒェンのうち外来語がこれより少ないのは 101 篇のうち 14 篇であり、伝説のうち外来語がこれより多いのは 293 篇のうち 68 篇で、いずれもそれなりに低い割合と言える。メルヒェンと伝説のデータ量が異なるため、単純に正解率を上げたい場合は 2.5%を境目とし、 $(43+25) / (101+293)$ で約 17.3%が間違い（約 82.7%が正解）となる。これは全てを伝説に分類したときの約 74.5%より高いため、外来語の割合は分類の基準として有効だと言える。ただし、こちらについても後述の HanTa と scikit-learn を用いて分類した場合より正解率は低いいため、やはり複数の品詞を用いて機械学習を行った方がより優れた分類を行えることが分かる。

機械学習は後述のとおり、複数の変数を用いて高次元で判定できるところに一つの長所があるが、アルゴリズムによってはその際にどの変数を重視するかの重みを付けられる（その重みの判断も機械がやってくれる）。外来語の割合は、外れ値が判定を狂わせる面はあるものの、それなりに重視しうる変数であると考えられる。

3-2 その他、メルヒェンと伝説で平均値に差がある品詞について

本論文は機械学習による分類を主題としているが、以上の説明からも分かるとおり、品詞分解した数値を文学者の目で精査するのみでも作品解釈の助けとすることは可能である。また、より多くの作品を対象とする場合には、機械学習のための説明変数をあらかじめ絞るのが有効な場合もある（変数が増えると、アルゴリズムによっては指数的に計算量が増えるためである）。このような理由から、本節では注目すべき品詞を選択し、それらについて考察する。

品詞の選択にあたって 2 点を重視した。1 点は、当然ながら M1 と S1 で平均値が大きく異なる品詞であること、もう 1 点は割合そのものの値が低すぎないことである。後者については、仮に平均値に数倍の差があったとしても少ない作品が大きな影響を与えている場合があるからである。以下、この 2 点に当てはまる品詞および単語数・語彙数について箇条書きで記す。

一、単語数および語彙数は、メルヒェンの方が 3～5 倍程度多い。これは、メルヒェンが物語として起承転結を備えていることが多いのに対し、伝説は事実の概略を述べたものが多いからである。ただし、「バイエルンのアーデルガ公伝説」(2438 語、829 単語)のように、M1 の平均値 (1507.86 語、467.75 単語) を大きく上回る S1 の伝説も存在する⁽⁴⁾。

二、普通名詞と固有名詞

固有名詞だけでなく、普通名詞も伝説の方が多。伝説の方が出来事を淡々と述べ

ているからとも想像されるが、理由ははっきりしない。これと同程度、メルヒェンに多く見られるのが外来語であるから、明晰な言葉で語られる伝説に対し、しばしば魔法のような不思議な言葉が用いられるメルヒェンにおいて、それらが名詞ではなく外来語に分類されているということかもしれない。

三、付加形容詞

伝説の方がやや多い。メルヒェンの方が修辞が多く思われるだけにやや意外な数値だが、「付加」するための名詞の割合が伝説の方が大きいことによるとも思われる。なお、叙述形容詞の割合に大きな差はない。

四、副詞

メルヒェンの方がやや多い。物語が展開していくために文の間に挟まるからとも思われる。

五、前置詞、冠詞

いずれも伝説の方が多い。付加形容詞と同じく、前置詞と冠詞は名詞を必要とすることから、名詞の割合の差として説明がつくと思われる。

六、人称代名詞

メルヒェンの方が多い。単語数の箇所ですべたように個々の物語がメルヒェンの方が長いので、すでに登場した人物について人称代名詞で言及する機会が多いと考えられる。

七、動詞における定形と完了形

HanTa は通常の動詞と状態を表すコピュラ動詞を区別してそれぞれの比率を示すが、いずれについてもメルヒェンの方が定形（現在形や過去形）が多く、伝説の方が完了形が多い。伝説の方がすでに起こったことを語る側面があるとも言えるが、あるいはメルヒェンの方が、会話文など現在形が用いられる箇所が多いのかもしれない。

以上のように、メルヒェンと伝説の違いは複数の代表的な品詞において割合の差として見られる。個々の物語ではなく、作品集としてのメルヒェン集と伝説集の違いを考察するのであれば、これらのうち幾つかの比率を用いれば十分に判定可能と思われる。これは、それぞれが作品集となることで、特殊なメルヒェン・伝説が持つ外れ値の影響が小さくなるからである。

本論文のように個々の多様な作品について分類を試みる場合、機械が分類できる確率を上げるために、多くの説明変数を渡して、1つの変数に外れ値が混ざることの悪影響を減らすのが有効である。以下、HanTa の品詞割合を全て scikit-learn に委ね、実装されているアルゴリズムによってどれほどの正解率で分類できるかを見ていき、また分類基準につ

いて考察する⁽⁵⁾。

4. 機械学習とその分類の正解率

Python には機械学習のためのライブラリが多数ある。ここでは scikit-learn を用いて 5 つのアルゴリズムに訓練用データとラベルを教え、学習成果をテスト用データで検証する。

4-1 決定木

「はい／いいえ」を選択して矢印を辿っていくと、どれに当てはまるかを示してくれる図は日常でもよく見る。機械がこれと同じやり方を用いるのを（逆さ向きの木のようなため）決定木と呼ぶ。分岐の最大の深さを指示することができ、分岐を細かくするほど正解率は上がりそうだが、いわゆる過学習⁽⁶⁾ が起こるので、今回は深さ 5 とした。

scikit-learn を用いたコード記述はどのアルゴリズムもほぼ同じで、それぞれのアルゴリズムモデルのインスタンス⁽⁷⁾ を作り、その fit 関数に訓練用データと正解のラベルを与えると学習が完了する。次に predict 関数にテスト用データを与えると、アルゴリズムに沿って分類してくれるので、汎用の accuracy_score 関数で正解ラベルと比較し、正解率を確かめれば良い。

決定木を深さ 5 で用いた結果、正解率は約 90.3%であった。決定木のメリットとして、どのように分類が行われたか分かりやすいことが挙げられる。次の図 1 は、図示のために深さ 3 までに絞った決定木の分岐表である（なお、深さ 3 だと正解率は 90.1%であった）。

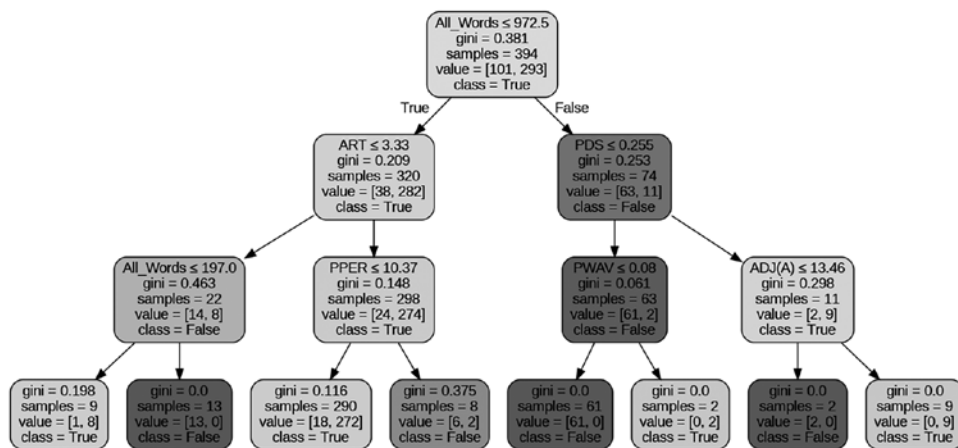


図 1

最初に単語数が972.5より多いか少ないかで分け、少ない場合には冠詞（ART）の割合が3.33より多いかどうかで、単語数が多い場合には指示代名詞（PDS）の割合が0.255より多いかどうかで分け、というように続いている。人称代名詞（PPER）、疑問副詞（PWAV）、付加形容詞（ADJ（A））の割合も判断基準に用いている。このような可視化しやすいアルゴリズムは、人間の判断のヒントにもなる。決定木は必ずしも精度の高くないアルゴリズムだが、この点でなお価値を持つ。

4-2 ランダムフォレスト

フォレストと付いている通り、決定木の仲間である。決定木は過学習が起こりやすいため、複数の決定木のモデルを作り、その中の多数決で決定する。理屈としては原始的だが、1つのアルゴリズムの名称を与えられているように有力な方法である。多数決を取るための決定木の数は引数として指定できる（調整してみたが、デフォルトの100とした）。約94.1%とかなり高い正解率を示した。

4-3 k近傍法

これは判定のためのルールを作るのではなく、説明変数分の次元の空間を用意してそこに訓練用データを置き、分類したいデータと近い位置にあるk個のデータのラベルの多数決で決めるという方法である。今回は57次元なのでイメージしづらいが、点と点の距離はそれぞれの次元での差を二乗して合計し、平方根を計算すれば良い。なお、変数同士の影響力を整えるためにデータの標準化をほどこすと良い。

k近傍法による正解率はk=10としたとき、約91.3%であった。今回試した5つのアルゴリズムでは素朴な決定木に続き低い結果だったが、それでも90%は超えており、どのアルゴリズムでもこの程度は正解できることが分かった。

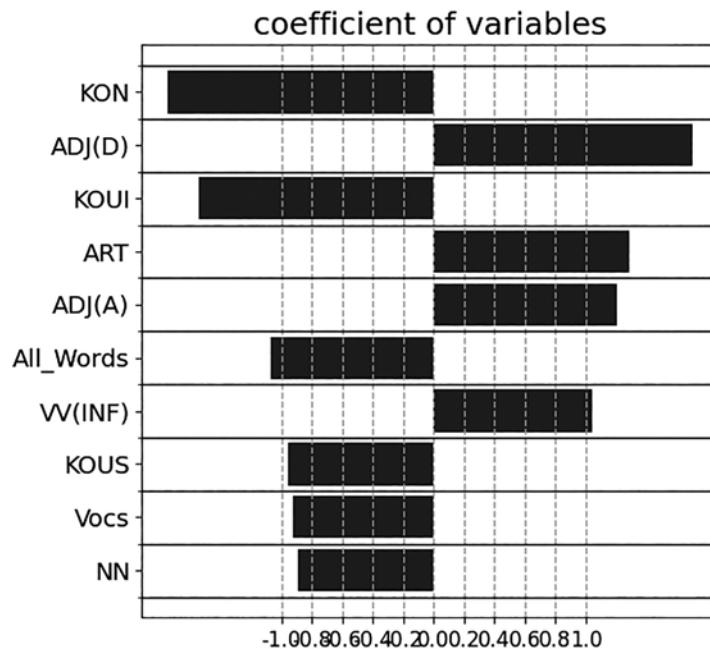
4-4 Linear SVM

データを2分する際に、空間上で最もマージン（分割する線とそれによって分割されるデータの距離）が大きい線引きを試みるアルゴリズムである。もちろん、今回のようにデータが完全に二分していない場合には、そもそも完全に分かれる線は存在しない。ある程度の許容値と試行回数を定めた上で、距離が最大になる線の傾きを決めることになる。強力なアルゴリズムとして知られ、今回も約93.1%の確率でテスト用データを分類してみた。

4-5 ロジスティック回帰

ロジスティック回帰は、それぞれの説明変数に重みとしての係数を掛けて最も分類が成功する方程式を作り出したのち、それによって生じた目的変数をロジスティック関数によって0から1に収めることで、質的データ（たとえばメルヒェンと伝説は連続するわけではないため質的データである）への分類を可能とする手法である。今回のように2項に分類する場合は、0.5を境にしてどちらか判断する。

ロジスティック回帰も有力な分類器として知られ、今回のテストでは約93.9%と高い正解率を示した。ロジスティック回帰については、インスタンスの訓練後にそれぞれの説明変数にかけた係数、つまりどの程度その変数を重視したかの重みを表示することができる。係数の大きさを比較する場合には、データの標準化が必要である。



グラフ 5

上掲のグラフ5は、説明変数への係数の絶対値が大きい順に10種を並べたものである。ラベルとしては0がメルヒェン、1が伝説であるため、マイナスの向きの品詞等はメルヒェンに特有の説明変数、プラスの向きの品詞等は伝説に特有の説明変数である。アルファベットは順にKON（並列接続詞）、ADJ（D）（叙述形容詞）、KOUJ（zu不定詞を取る接続詞）、ART（冠詞）、ADJ（A）（付加形容詞）、All_Words（単語数）、VV（INF）

(動詞不定形)、KOUS (従属接続詞)、Vocs (語彙数)、NN (普通名詞)であるから、接続詞・単語数・語彙数・普通名詞の数などはメルヒェンと判定するための変数となり、形容詞、冠詞、動詞不定形などは伝説と判定するための変数となっている。3-2で私が記した判断と一致するところが多いが、普通名詞については異なる判断をしている。これは、単に普通名詞について判断が分かれているのか、あるいは他の変数と組み合わせたときに何らかの補正をするのに普通名詞が適しているのか、理由は定かではない。

4-6 多数決

アルゴリズムそのものではないが、複数のアルゴリズムを組み合わせることでそれぞれの弱点を補うことができる多数決という方法がある。多数決 (アンサンブル分類と呼ばれる) は、scikit-learn にもあらかじめ実装されており、上記のランダムフォレストもその一部である。ここでは、正解率の高かった3つのアルゴリズム (ランダムフォレスト、Linear SVM、ロジスティック回帰) の多数決によって分類結果を得る。

複数の分類器による多数決には2つ方法があり、単純に投票数による多数決にする方法と、それぞれのアルゴリズムの判定における所属確率 (つまり、判断についての「自信」) を利用して、平均確率が上回った方に分類する方法がある。後者の方が緻密な手法であるため有用に思われるが、精度の低い分類器が確信を持って間違えた際には、その判断に全体が引きずられるという欠点もある。今回は単純な多数決による判定では約 94.1%、平均確率による多数決による判定では 94.4%と、わずかに後者が上回った。これが今回の5つのアルゴリズムを用いたうえでの最も高い判定確率である。

本章では、5つのアルゴリズムとそれによる多数決について検証した。いずれも 90.3 ~ 94.4%と高い正解率を示し、物語の内容に踏み込まなくとも、HanTa を用いた形式的な数値だけから、scikit-learn のシンプルなコードを用いることで、かなりの確率でメルヒェンと伝説を区別できることが分かった。

5. 分類に失敗した作品についての考察

最後に、機械学習が分類に失敗した作品についてその理由を考察する。前章では正解率のみに着目したが、それぞれの分類器はテスト用データの各物語をどちらに分類したかも示してくれるため、どの作品を分類ミスしたかは容易に分かる。ただし、5つの分類器それぞれの結果について考察するのは重複もあり、煩雑であるため、ここでは平均確率による多数決の結果を元に考察を進める。

5-1 メルヒェンであるのに伝説に分類された作品について

テスト用データ M2 には 100 話が収められている。そのうち、判定に失敗したのは「ならずもの」「歌う骨」「狼と名付けをたのんだ奥さま」「ハンスの嫁取り」「天のからさお」「下男」「神様の動物と悪魔の動物」「星の銀貨」「親すずめと四羽の子すずめ」「ガラスの棺」「天国へ行った水のみ百姓」の 11 話である⁽⁸⁾。なお、伝説をメルヒェンに分類するよりも誤答の確率が高くなっているのは、訓練用データにおいて伝説の方が数が多いため、正解率を上げようとすれば伝説と判断した方が良いことに起因すると考えられる。

まず、当該作品の単語数・語彙数および品詞割合の表を見て、どの説明変数のために伝説に分類されたかを検討する。今回は複数の分類器で確率に基づく多数決を取っているため、これまで見てきたロジスティック回帰や決定木の基準を元に判断理由を推測する。

まず、それぞれの説明変数でソートしてみたが、はっきりと 11 篇全てが多くまたは少なく分類される変数はなかった。たとえば単語数で見ると 10 篇が少ない順に 37 篇目までに収まるが、「ガラスの棺」は 83 篇目である。あるいは、付加形容詞の割合で見ると、上位 6 篇に 5 篇入っているが、「ならずもの」は下から数えて 6 篇目である。したがって、機械が一つの変数に拠ったわけではなく、変数を複雑に組み合わせて判断した結果であることが読み取れる。なお、「ガラスの棺」は前置詞の割合が 100 篇で一番多く、付加形容詞が「ハンスの嫁取り」の次に多い。また、「ならずもの」は関係代名詞とコピュラ動詞不定形が一番多い。この辺りの極端な数値がこれらの作品を伝説に分類させたと思われる。

上記の 11 篇について、人間の目で伝説に分類する可能性はかなり低い。唯一、「神様の動物と悪魔の動物」はキリスト教をベースとし、「コンスタンティノーブルの教会」という具体的な地名が出てくるのでやや伝説的だが、仮にこの「コンスタンティノーブル」を具体的な場として捉えるのであれば、そもそもドイツの民間伝承を集めた『グリム童話』と『ドイツ伝説集』に含むとも思われない。このことから、機械は人間には分類可能なものを誤って分類していることが分かる。ただし、こうした物語について判断を誤ったにもかかわらず、おおよそ 10 話のうち 9 話は正しく分類できているところには、人間に再現しがたい賢さを機械に認めるべきかもしれない。

5-2 伝説であるのにメルヒェンに分類された作品について

テスト用データ S2 の 292 話のうち、メルヒェンに分類されたのは、「ホッラ婦人と農夫」「地中の小人と羊飼いの少年」「粉屋の家の精」「ヒンツェルマン」「悪魔の蹄鉄」「パンの靴」「境界を決める駆けっこ」「小人と不思議な花」「三人の宝掘り」「白鳥を伴う騎士」「パンと塩を祝福する神」の 11 話である。判別が難しい例については、機械は伝説に

分類するはずであるから、この11話についてはより根拠があってメルヒェンに分類していると考えられる。また、土地ものの上巻に9話、歴史ものの下巻には2話と、土地ものに判別ミスが集中していた。

この11篇について、いずれかの説明変数でまとまったグループを成すことはなかった。メルヒェンよりも作品数が多いため、複数の作品が並ぶ確率は低くなる。たとえば前置詞の比率でソートして、11篇は小さい順に122篇の内に収まるため、前置詞の少なさはメルヒェンに分類されやすい原因と考えられるが、前置詞の比率が最も少ない8篇は伝説と判定されている。単語数と語彙数の多さでは1位が「ヒンツェルマン」、2位が「白鳥を伴う騎士」となり、この2篇は両者の多さでメルヒェンと判定されているようだが、「三人の宝掘り」はそれぞれ少ない順に23位（単語数）、25位（語彙数）である。メルヒェン同様に、少数の説明変数で導き出しているわけではなく、複数の変数を組み合わせて判定しているようである。「三人の宝掘り」は冠詞の多さで3位、数詞で7位、付加疑問代名詞で2位となるやや独特な品詞割合の作品ではあるが、それぞれの割合についてこの作品を上回る各物語が伝説に分類されている。

次に、メルヒェンに分類されたそれぞれの伝説の内容を見て、人間の判断でもメルヒェン的であるかを確認する。まず、いかにも伝説らしいものから見ていく。「地中の小人と羊飼いの少年」は冒頭で年代と場所が指定されているため、グリム兄弟の基準に照らして判断すれば、明らかに伝説である。「粉屋の家の精」、「悪魔の蹄鉄」、「小人と不思議な花」、「三人の宝掘り」は、それぞれ地名が複数回登場するため、それを根拠に伝説と判断できる。ただし、精霊や悪魔や小人が出てきて教訓を学ばされるところはメルヒェン的でもある。「ヒンツェルマン」は地名も年も指定され、登場人物も固有名が明示されているため、むしろ歴史に近いが、精霊が理解不能な力を示すことで伝説に分類される物語である。物語の長さからメルヒェンに分類されたと思われる。「パンの靴」は固有名詞のない珍しい伝説で、不思議なことも起こり、子供の理解が難しい物語でもないため、内容的にはメルヒェンに分類されてもおかしくない。逸話的で長さが短いこと、同様にパンの靴が出てきて地名も明示される「白パンの靴」の類話であることからグリム兄弟は伝説に分類したと思われる。同様に「パンと塩を祝福する神」は、冒頭に「ドイツ人」と出てくるのみで土地も時代も明記されず、不思議な出来事が起こる物語であるため、グリム兄弟の判断基準でもメルヒェンに分類しうる。「境界を決める駆けっこ」は、村の境界を定めるために、それぞれの村の鶏が啼いてから代表者が競走をしたという物語で、そもそも理解不能なことが起こらない。地名が登場すること、またおそらくはグリム兄弟の判断として、土地の境界を決めるのに競走で決めるのはありえないことから伝説に分類したと思われるが、「理解不能なものと常に混ぜ合わせている」という彼らの定義を厳密に適用するなら伝説

ではなく、もちろんメルヒェンには当たらない。

興味深いのは、メルヒェンに類話がある「ホッラ婦人と農夫」と「白鳥の騎士」が誤分類されたことである。「白鳥を伴う騎士」は歴史ものの伝説として地名や人名を多く伴い、明らかな伝説であるが、兄弟が白鳥になって飛んでいくという筋書き自体はメルヒェン「六話の白鳥」と共通する。「ホッラ婦人と農夫」については、上述のとおり、幾つもある“ホッラ婦人もの”についてグリム兄弟が境界上の物語であることを認めている。『ドイツ伝説集』に収められた“ホッラ婦人もの”のうち、「ホッラ婦人と農夫」以外の4篇については場所が明示されており、人間の判断でも伝説に分類できる。これに対し「ホッラ婦人と農夫」には固有名詞が「ホッラ婦人」しか用いられていない。この作品を伝説に分類する根拠があるとすれば、「パンの靴」や「パンと塩を祝福する神」同様、話が短いため物語性が薄く、逸話のようだという点に尽きる。このため、「ホッラ婦人と農夫」をメルヒェンに分類した機械の判断は、「場所や、史実となった名前に結びついている」ことを伝説の基準とするグリム兄弟あるいは人間の目から見てもそれほど違和感がない。

以上のように、多様な作品が誤分類されたものの、メルヒェンを伝説に誤分類した例に比べれば、人間の目からも納得する例が多いように思われる。これが、より慎重な基準にしたがって機械が分類したことを理由とするならば、機械の慎重な判断は（人間と違う判断基準にもかかわらず）、人間の境界判断と連動していると考えられる。

おわりに

本論文は、HanTaを用いて品詞分解し、それをもとに機械学習することによって、物語の内容に踏み込まなくとも95%近い確率でメルヒェンか伝説かを判定できることを確かめた。HanTaが方言に対応できていないことで、最も重要と思われた品詞である固有名詞、そして外来語の割合を実際ほど活用できないという問題点も明らかになったが、多数の品詞の活用でそれを補う、高次元処理の長所も示すことができ、ツールの検証として有意義なものになったものと考ええる。

本研究の成果を元に、メルヒェンと伝説の品詞割合についてより考察を深めることも可能であるし、グリム兄弟以外の編者によるメルヒェン集や、ムゼーウスなどの作家性の強いメルヒェン、さらにティークラの創作メルヒェンとの比較を進めることも有意義である。今後の研究課題としたい。

注

- (1) HanTaについて使用したバージョンの変化に伴い、品詞分類が変化している。表記の変更

- (たとえば VVFIN から VV (FIN)) は形式的なものだが、新たに NNA (形容詞の名詞化) と NNI (動詞の不定形の名詞化) が ver. 1.1.0 から追加された。
- (2) 以下、ドイツ語文献の引用は全て、参考文献に載せた既訳 (とりわけ鍛冶・桜沢訳) を参照し、文脈に合わせて筆者が訳した。
- (3) とりわけ「漁師とおかみ」と「もみの木」は、ロマン主義画家のフィリップ・オットー・ルンゲが提供し、グリム兄弟がその方言も含めて称賛したメルヒェンである。
- (4) 後述のとおり、S2 の「ヒンツェルマン」(6711 語、1869 単語)、「白鳥を伴う騎士」(3685 語、1101 単語)、また「カール王」(2540 語、911 単語) などメルヒェンの平均値を大きく上回る伝説である。
- (5) 以下の正解率については、メルヒェンと伝説の二択ではあるが、訓練用およびテスト用データのメルヒェンと伝説の割合に差があるため、基準は 50%ではなく、全てを伝説と考えたときの約 74.5%になることに注意が必要である。データの比率がより極端な場合は「不均衡データ」として補正を試みるが、本研究では特に補正処理を加えていない。
- (6) 過学習とは、与えられた訓練用データのみにも最適化された細かい分類規則を機械が学んだ結果、汎用性を失う現象のことである。極端な例を上げれば、物語それぞれの単語数を暗記し、「1034 字ならメルヒェン」「502 字なら伝説」というように条件を作れば 100%の分類規則が作れるが、これはもちろんテスト用データを正しく分類できない。
- (7) インスタンスとは、あるモデルをそれぞれのプログラムにおいて使用できるように実体化した (メモリを割り当てた) ものである。喩えるなら、あるモデルのテレビは規格として統一されているが、各家庭で見られるために個々に製造され、各家庭の視聴スタイルを覚えていくように。このようにして作られた個体をプログラム上でインスタンスと呼ぶ。
- (8) あくまで、テスト用データからの 11 話であるため、この 11 話が『グリム童話』全体のなかで特に伝説に近いと示されたわけではないことには注意が必要である。

参考文献

Projekt Gutenberg-DE の各テキスト

Grimm Brothers. (Eds.). (n. d.). *Deutsche Sagen*. Retrieved August 20, 2023, from <https://www.projekt-gutenberg.org/grimm/sagen/sagen.html>

_____. (n. d.). *Kinder- und Hausmärchen Vollständige Ausgabe*. Retrieved August 20, 2023, from <https://www.projekt-gutenberg.org/grimm/khmaerch/khmaerch.html>

Grimm Brothers (Eds.). (1816-1818). *Deutsche Sagen*. Nikolaischen Buchhandlung.

_____. (翻訳) 吉田孝夫訳 (2021 年)『グリム ドイツ伝説集』八坂書房。

_____. (翻訳) 鍛冶哲郎、桜沢正勝訳 (2022 年)『グリム ドイツ伝説集 新訳版』鳥影社。

Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2* (3rd ed.). Packt Publishing.

_____. (翻訳) 福島真太郎監訳 (2020)『Python 機械学習プログラミング 達人データサイエンティストによる理論と実践 第 3 版』インプレス。

Wartena, Christian. (2019). "A Probabilistic Morphology Model for German Lemmatization." In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, edited by the Chair of Computational Corpus Linguistics, 40-49.