

言葉を素材にした分析論

文章分析の基本機能

(株)シードウィン

分析対象の文字数

分析対象

文字で表現されているすべてのデータを対象とする。
単語群の集合、文節の集合、1センテンス(箇条書き)、文章など。

1文字だけの単語は分析できない。複数の単語群が集まって分析できる。1データの文字数が少なくても、アンケート、メモなども複数のデータが集まれば、分析対象となる。

韻文(短歌、発句など)は、分析基準が異なってくる。ここで表されている分析方法では分析できない。古文は分析対象となるが、現段階では精度は低い。現代文を対象とする。古文は異なる言語として扱う方が適切である。

以下は1文章を分析対象として、1文章のみで分析結果を求める場合。

50字以上100字未満

単語分解と分野分類のみの分析となる。

100字以上400字未満

100字、200字、300字の分析基準があり、それぞれの基準で分析できる。400字以上の分析では35項目の分析値が求められているが、400字未満では、求められない分析がある。

400字以上600字未満

400字の分析基準がある。この基準での分析可能文字数は350字～600字である。適正な分析結果が求められる。400字未満の場合、文章の表現技術によっては求められない分析値がある。

600字以上2000字未満

ここで表されている機能はすべて分析結果として抽出できる。

2000字以上

1テーマで表されている文字数はおよそ2000字以内である。2000字以上であっても分析に問題はないが、テーマが複数存在する確率が高い。この文字数を超える場合は、テーマ単位に分割する必要がある。

基本分析機能 1

A 基本分解分析機能

基本分析は一定の文字量の文章を必ず分解し、分析をする項目を挙げている。すべての分析結果には、ここで挙げられている分析内容はすべて実施されている。文章を分解するために辞書が必要となる。体言辞書、用言辞書、反意語・類語辞書などがある。カタカナ、アルファベットで登録されていない単語群は自動で分解される。

1

文章を単語単位で品詞設定

a

最小品詞単位で分解する。複合語は、複合化される前の単語と複合化された単語群に区別される。複合化されて意味が変化する場合は複合語で1つと判断されている。単語別、品詞別、分野別、活用の一部の意味別にカウントする。

2

文章中単語の意味の重さ

a

すべての各単語はカウントされ、文章全体の文字量に合わせてカウント比率が求められる。基準が文章文字数になっているので単語の文字数を計算し、他の文章との比較ができる数値に置き直す。相対的比率を分析値に置き換える。この値を**単語重量値**と言い、文章中で作者による単語の意味の重さとする。

3

基準単語への変換

a

表現の仕方で、短縮形などを標準の表現に置き換えている。動詞、形容詞、助動詞は、終止形に置きかえてから分析を行う。データは元文と標準にした単語群で構成された2つを持つ。

a

1文章単体で分析できる項目、複数文章も同じ

g

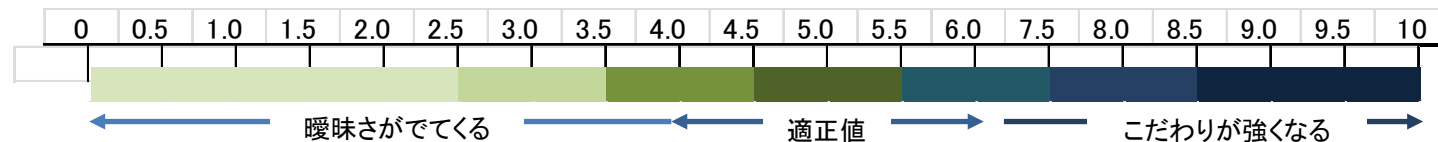
複数文章で分析できる項目

基本分析機能2

4 他文章比較用分析値抽出

a

35項目の分析値がある。正規分布している分析値を16項目取り出し、常に表示できる値として活用している。表される分析値のすべては、異なる文章、異なる文字量の文章などと比較しても同じ意味で分析値を比較できる値にしている。各分析値の意味も同様の扱いになっている。分析値の基準は5.0であり、この値より大きくなるとこだわりが出てきて、小さくなると曖昧になっていると判断できる。



5 同意語・反意語の抽出

a

分析時に使用された単語群に対して、同じ文章中で同意語と反意語が抽出されている。これらは、伝達意識を計測したり、他データの検索要因として使われる。

6 習慣使用単語の抽出

a

1つの文章で、よく使われている連体詞、接続詞、副詞、感嘆詞、などが抽出されている。

7 キーワードの抽出

a

第1キーワード群

第1補完単語群

一頻度単語群

右のように、1文章に表された単語を7つのブロックに分類している。第1から第3のキーワード群が適正に書かれている文章で4%程

第2キーワード群

第2補完単語群

第3キーワード群

第3補完単語群

度、1頻度単語群が80%前後の比率で構成されている。この構成比で、文章表現の適正さが検証できる。趣旨、キーセンテンスもこの構成から計算される。

8

キーセンテンスの抽出

a

文章中での単語の意味の重さ(=単語重量値 以下 W_g と表す)を用いて1センテンス重量値を現す。1センテンス重量値は感嘆詞、連体詞、接続詞を除く体言と用言の W_g 合計とこれらの単語数から平均を求める。さらに、1センテンス内にキーワードが含まれているセンテンスで、1センテンス重量値の大きいものをキーセンテンスとする。

⇒単語重量値の説明は『文章分析<文道>の分析方法』を参照

9

思考&テーマ分野の抽出

a

体言辞書の単語の1つ1つに分野が設定されている。文章で使われた単語群で辞書に分野が設定されている単語の出現数とそれらの W_g から分野が計算される。1分野の固定ではなく、複数分野の比率で表される。

10

近似文への変換

a

センテンスから付属語(助詞、助動詞)を除いたセンテンスを近似文と言う。近似文は他のセンテンスの同一性の発見、他言語への変換に用いられる。

11

テーマ分野自動設定

a

特定分野について、自動設定されている。分野辞書を用いるのではなく、政治、経済、国際、金融、社会、事件についてのみ自動設定される。これらは複数分野の特定ではなく、1分野特定に用いられる。

キーワード群の相関 1

B キーワード群相関抽出

1文章の趣旨は『日本語の構造』で示したように使われている単語群の20%で表現されている。その中心となるのがキーワード群である。

『基本分析機能2』の「キーワードの抽出」で表しているように単語群を7段階に分類している。この分類は構成比で分類するのではなく、単語重量値の数で分類する。従って1文章で表現されている内容に主題が存在せず、内容が分散しているときは、キーワード群は検出されない。

複数文章の相関・相反を分析するとき、Wgとキーワード群を用いて関係が算出される。縦軸に複数文章の共通する全体のキーワード群をWgの降順で並べられ、横軸に各文章を並べ、縦軸と共有した単語のWgを並べる。

= MR活動に対する医師アンケート =

単語	Wg	カウント数	Wg平均	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
接待	3.7	44	4.0	● 3.7	● 8.4	● 5.4	● 6.1	● 2.4	● 2.8	● 3.5	● 5.8	● 4.7	● 3.7	● 4.2	● 3.2	● 3.2	● 2.7	● 2.6	● 2.8
良い	2.3	40	2.4	● 2.3		● 3.3		● 1.6	● 2.1	● 1.0	● 2.1	● 3.9	● 2.5	● 2.2	● 1.9	● 2.6	● 2.0	● 1.8	● 3.5
MR	1.2	30	1.4			● 1.7		● 1.2		● 2.1	● 2.1	● 1.4	● 2.1	● 1.6	● 1.1	● 1.9	● 0.9	● 1.0	● 2.1
受ける	0.9	21	1.3					● 1.6				● 1.1	● 1.3		● 0.5	● 0.6	● 1.3		● 1.0
必要	0.8	20	1.1										● 0.6		● 0.9	● 0.7	● 0.9		● 1.3
規制	0.7	17	1.2					● 1.2	● 2.1					● 1.1	● 0.6	● 0.9		● 1.3	● 1.4
会社	0.8	16	1.1			● 2.5				● 1.0			● 0.9		● 1.1	● 0.6	● 0.9	● 1.3	
情報	0.7	16	1.0							● 1.4		● 1.9	● 0.9	● 1.3	● 0.5				

上表は、MRの接待活動に対する医師アンケートでキーワード群の相関を抽出した一部である。横軸にAから始めた45名があり、縦軸に45名全員のキーワードを並べている。最初のWgは、縦軸の値で、Wg平均は横軸の数値の平均である。カウント数は縦軸の単語をキーワードとした人数を現している。

上表をワードクロスとしている。

1

共通認識を単語で抽出

g

上表のカウント数が大きいほど、共通意識の強い単語となる。但し、縦軸のWgが高くても、個別データによっては、低いものもある。全体意識と、個別データの意識の差が現れる。

キーワード群の相関2

2

無認識下での共通認識抽出

g

『キーワード群の相関1』の表で縦軸にワードが現れても、横軸でキーワード群に当てはまらない単語群が現れる。個別データとしては意識されていないが、全体ではキーワード群になっている。このような単語群を無意識下での共通認識となる。1個人の3文章以上を分析し、ワードクロスを表したとき、個人の無意識の意識単語、つまり暗黙知が現れる。

3

伝達率の測定

g

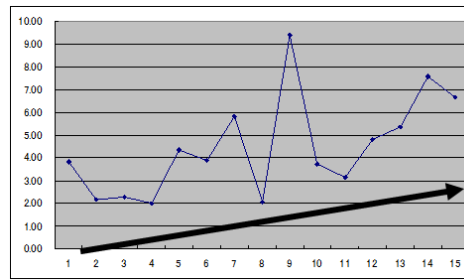
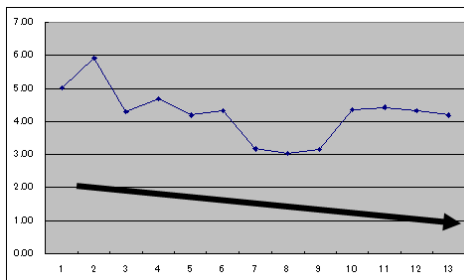
伝達したい内容があつて、その伝達されたはずの内容に伴って行動した事柄を記述(例えば報告書)した場合、ワードクロスで表されたところから、意識化、無意識化での単語群とそのWgが抽出される。

4

意識分野相関

g

キーワード群から、単語に伴った分野と比較して、意識分野が抽出される。『思考&テーマ分野の抽出』の発展形である。



左グラフは1人の文章を2年間継続して分析した時の2単語の変化である。縦軸はWgを表し、横軸は月数を表している。意識が強くなっている単語、次第に弱くなっている単語がある。弱くなっていく単語群、強くなっていく単語群で、意識変化が読み取れる。

る。複数の単語から分野が想定されるので意識される分野変化がつかめる。

C 分析の種類

AとBの分析技法、さらに別に基準となるデータを用いて、分析結果を導き出す。分析の概念には、次の6書類がある。

- ①.基準分析・・・基準となるターゲットを求めようとする分析、
- ②.比較分析・・・基準を比較しその関係を導き出そうとする分析、
- ③.絶対分析・・・データ単独での状態・性質を求める分析、
- ④.相対分析・・・複数データを単独で分析し、その相関を求めようとする分析、
- ⑤.直接分析・・・データの作成者を求めようとする分析、
- ⑥.間接分析・・・データの内容を調べようとする分析の6種類がある。

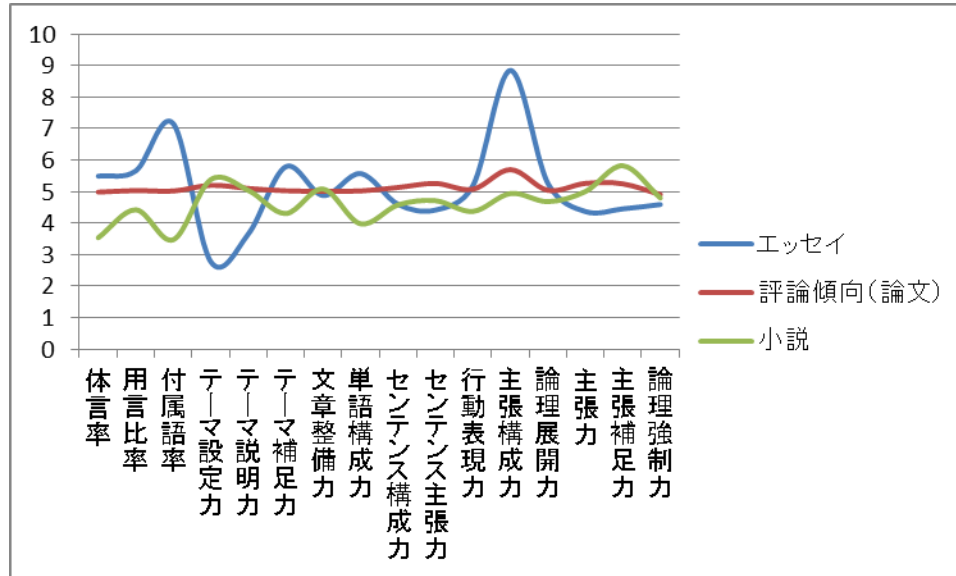
分析概念の名称は《文道》についてのみ使われている言葉で他の分析概念の表現とは異なっている場合がある。分析は①～⑥の組み合わせで行われている。『基本分析機能』で表した11項目は分析概念の絶対分析に当てはまる。『キーワード群の相関』で表した4項目は分析概念の相対分析になる。この項目「標準分析機能」は前述の2つの分析機能と、比較すべき基準を別にもって行われている。基本分解分析機能は、言葉で表現する基準が求められて、その基本とのギャップを測定して表しているとも言える。基本となるターゲットは言葉で表現する表現構造と単語の構成である。これを表現スタイルとしているが、表現に現れる感情は除外され、伝達のみを表している。『基本分析機能』は分析目的に応じて、別基準がそれぞれ存在する。

様々な測定 1

1

表現ジャンルの抽出

a



表現ジャンルは文学ジャンルと同じ分類方法であるが、論文、エッセイ、小説の3分類で表し、1文章の結論の有無で区分する。文学ジャンル別に、16項目の分析値が求めた。その分析値を折れ線グラフに表し、折れ線グラフの形状比較をした。その形状は、およそ3つに区分され、その区分が論文、エッセイ、小説形式とした。

《文道》では分析値を求めている。この分析値の基準は、新聞社説、一般論文、評論の数万の文章からターゲットを求めた。これを論文基準としている。指定がない限り、論文基準を分析値基準としている。論文基準、

エッセイ基準、小説基準、または職種別基準など、基準は目的に応じて設定できるようになっている。他の基準も論文基準と比較した値をだしているの、基準が異なっても比較検証ができる。

2

表現テクニカル測定

a

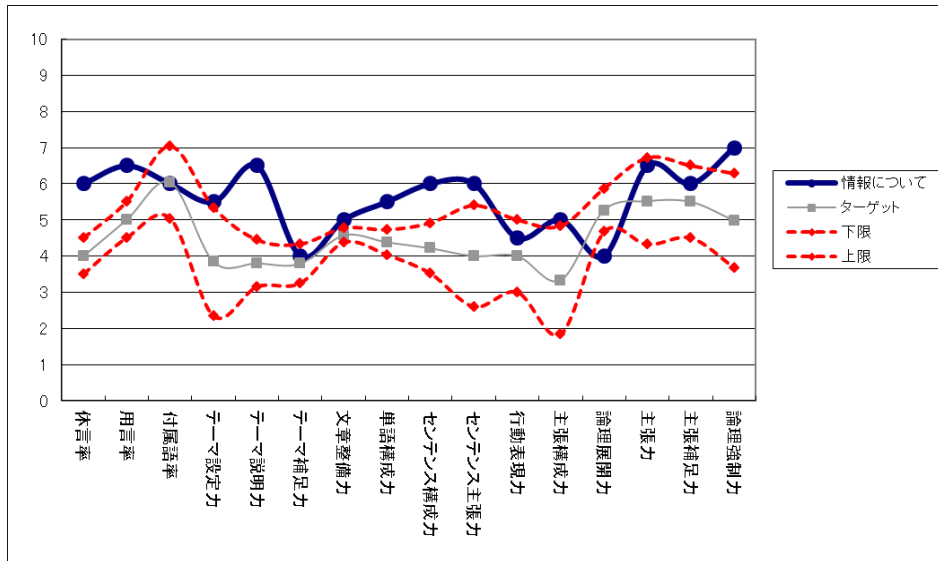
文章構造、文章中の使用単語群構成から、10項目の表現条件が上げられている。体言と用言の構成比、指示語、副詞、接続詞の使い方、論調の統一性、「こと」「もの」の使い方など。さらに、キーワード群の使用適正、1センテンスの文字量の適正などを測定している。

様々な測定2

3

表現好感度の測定

a



分析値を折れ線グラフ(左グラフ青線)にした。この形状から思考&表現スタイルを抽出する。このグラフを『文章の心電図』としている。折れ線の形状と、分析値範囲から抽出する。
折れ線形状は全体で11種類のコードで表され、形状分類だけで14,348,907通りがある。左グラフの赤破線の範囲が好感度が良いとされる範囲である。

2

思考&表現スタイルの測定

a

人材パターン項目	右上がり型	右下がり型	平行型	山型	谷型
対人性	惹付型	押し付け型	観察型	積極型	受け身型
論理性	列記型	語り部型	評論型	説明型	対応型
表現性	広がり型	まとまり型	伝達型		
強調性	吐き出し型	命令型	黙考型	カミ型	分散型
思考性	集中型	混迷型	整頓型		
主張性	説得型	論証型	まとめ型	論理型	押し付け型

上図の16項目の分析値が6ブロックに分けられている。

⇒「文章分析<<文道>>の分析方法」を参照

それぞれのブロックのグラフ形状から、左表のように分類され、思考・表現スタイルが分類されている。

様々な測定2

4

思考力の測定

a

⇒思考力分類は、人文組織工学人材論「人材能力について」を参照

思考力は11の能力で定義されている。その定義に従って分析値より計算される。《文道》での定義は心理学に従っているが、組織によって定義が異なる場合がある。その定義性が明確であれば、組織に応じた分析値が求められる。

目的別分析機能

1	入社エントリーの分析	a
2	昇格試験論文評価	a
3	研修効果測定	g
4	研修前効果設計測定	g
5	適性診断	a
6	組織診断	a
7	文章の作者の確定or特定	g
8	相性程度の測定	g
9	価値の抽出	g
10	現象からのニーズ抽出	g

アンケート分析

1	意識ターゲット・ベクトル抽出	g
2	意識カテゴリ分類	g
3	商品と市場ギャップ抽出	g
4	商品技術カテゴリマップ	g
5	商品群からの意識変化抽出	g
6	商品群相関調査	g
7	商品機能市場伝達率	g
8	意識ターゲット・ベクトル抽出	g
9	表現過不足からの伝達率	g
10	表現戦略要素の抽出	g

a

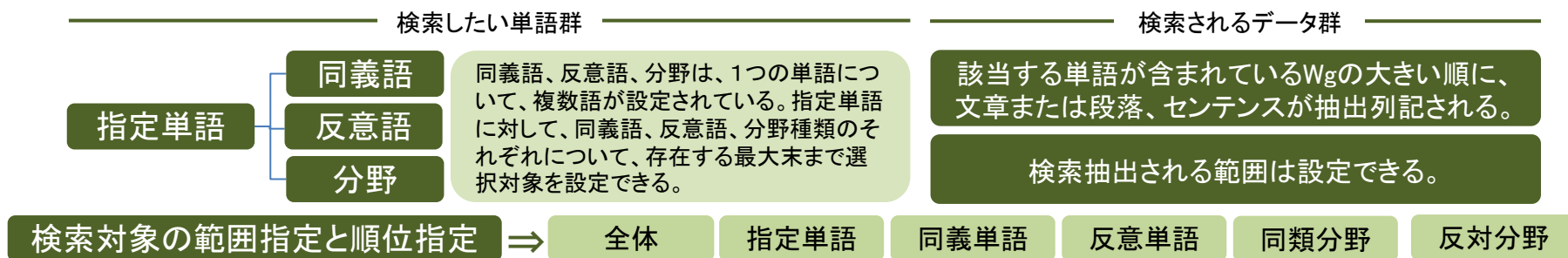
1文章単体で分析できる項目、複数文章も同じ

g

複数文章で分析できる項目

D 《文道》による検索機能

《文道》による検索は、単語重量値(以下 W_g と表す)およびキーワード群分類によって行われる。検索対象のデータ群は、分析処理された後、キーワード群、キーセンテンス群が抽出され、1文章のすべての単語に W_g が付けられている。検索対象は、適合する単語群の W_g の大きい順に検索され、表示される。単語群の中で分野コードが付いている単語がある場合は、一致または、類似する分野の文章群も、その文章群の中で一致する単語の W_g の大きい順に抽出される。



検索したいキーワードと同一単語を含むデータを検索するのであるが、検索対象となったデータで、検索したい単語群が、対象データの中でキーワードになっていなければ対象外となる。

1文章で検索対象ワードが1回しか使われていないデータは検索されない。1文章で1回しか使われない単語群は使用単語数の80%以上である。これに該当するデータは外される。キーワード群はおよそ5%以内であるから1/20に絞られる。2単語の積集合による検索になるとさらに絞り込まれる。

絞り込み係数(0~7段階)を設定すれば、絞り込みの程度を限定できる。0の場合は、該当単語が存在すれば抽出する。

他市場、他分野からの影響を抽出する場合、他分野への影響度を抽出するのに活用される。

検索機能2

1 複数単語からの検索

g

入力された単語の順に、対象文章全体から、上記の方法で検索抽出される。

2 1センテンスからの検索

g

検索するために入力されたセンテンスを単語に分解し、その単語を検索単語とする。意図する事柄をセンテンスにして、そのセンテンスを検索キーにできる。

3 文章からの検索

g

検索するために入力された文章を分析し、その中からキーワード群及び、キーセンテンス群を抽出する。それらを単語単位に置き換え、検사용単語とする。

4 分野からの検索

g

分野単語を検索単語として検索する。この時は、分野が一致、類似している文章群からの検索になる。

5 他データとの結合処理

g

右図のように、表現事前、表現事後、表現時等の限定条件や必然条件がある。行動目的に応じて、それら事項、事例をデータベースで用意し、表現された内容に対して事前指針、事後指針を提供する。

