

Webフォントの現在

飯塚 修平

2026/01/23 JEPAセミナー

自己紹介

飯塚 修平（いいつか しゅうへい）
@tushuhei

『Webフォント実践ガイド Google Fontsで
はじめる美しく機能的なタイポグラフィ』
（技術評論社） 著者

工学博士。外資系IT企業にて国際化を考慮したウェブサイト開発から機械学習システム構築まで、幅広く携わっている。



今日の内容

- 文字とフォントの基礎知識
- さまざまなフォント技術
- Google Fontsでできること
- 異体字の解剖
- BudouX: 機械学習とタイポグラフィ



A $\xrightarrow{\text{符号化}}$ 1000001
 $\xleftarrow{\text{復号}}$

normal diagonal-fractions stacked-fractions Unicode (U+00BD)
1/2 ½ ½ ½

She might
She might

文字とフォントの基礎知識

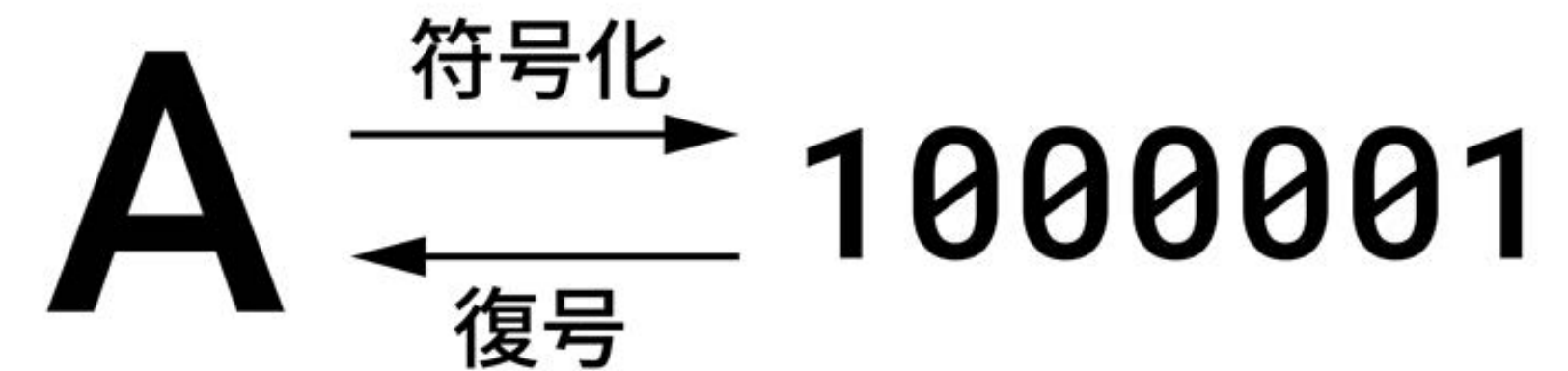
ac



コンピュータにおける文字の表現

符号化文字集合、Unicode、文字符号化方式

- 文字も二進数の値（コードポイント）で表される
→ **符号化文字集合**
- 世界中のあらゆる文字をカバーした符号化文字集合
→ **Unicode**
- コードポイントをバイト列に変換する方式（UTF-8やShift_JISなど）
→ **文字符号化方式**



文字とUnicodeコードポイント

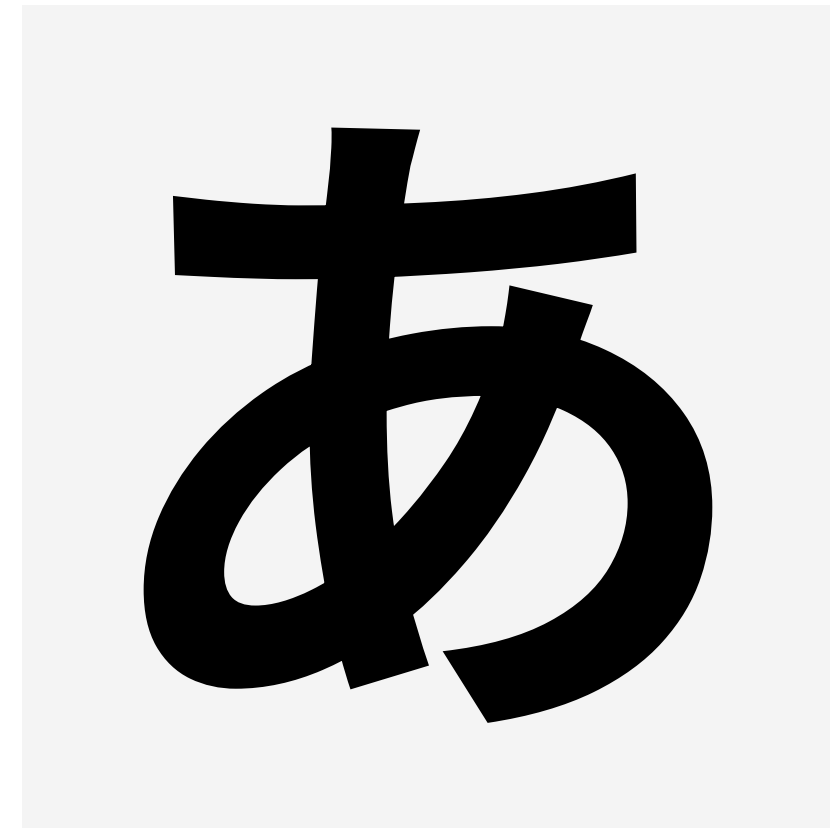
A	U+0041
あ	U+3042
亜	U+4E9C
🤔	U+1F914

文字が画面に表示されるには グリフ、書体、フォント

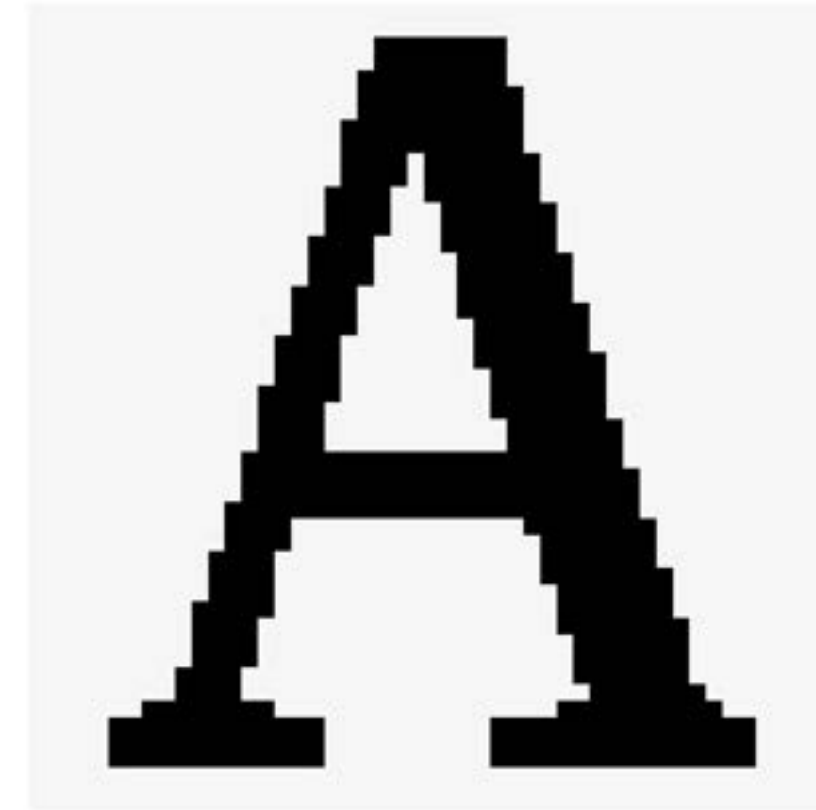
- 具体的な文字の形 → **グリフ**／**字形**
- **字体**: 点画で表される文字の骨組み
- 共通した特徴を持つグリフの集まり → **書体**
- グリフを定義するデジタルファイル → **フォント**



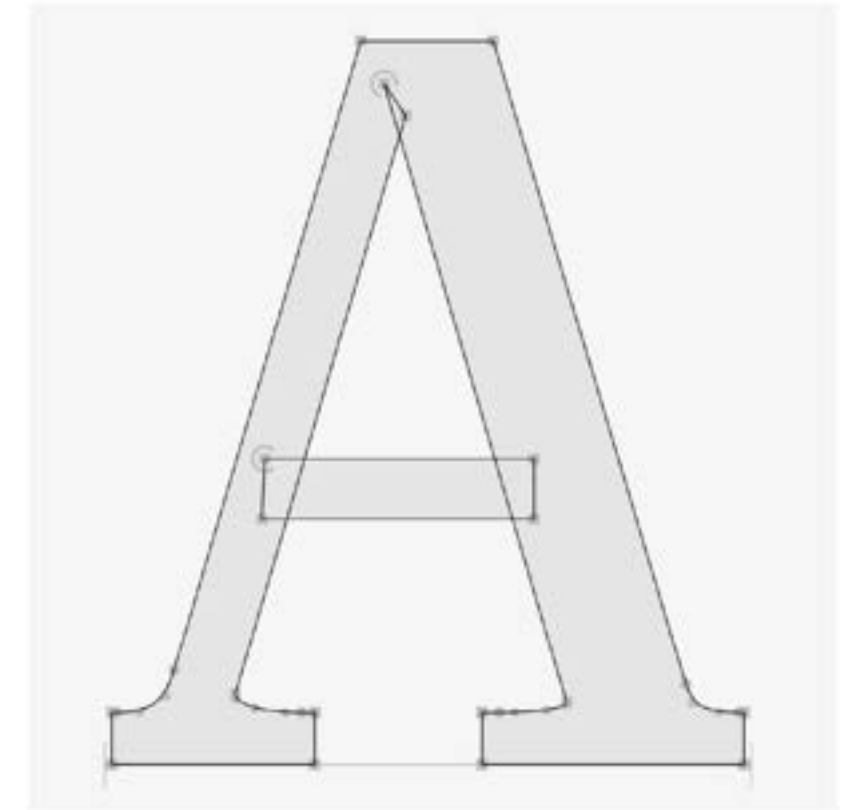
明朝体
Noto Serif JP



ゴシック体
Noto Sans JP



ビットマップフォント



アウトラインフォント

セリフ／ウロコ

- 文字のストロークの末端の装飾
→ **セリフ／ウロコ**
- セリフの有無で大きくセリフ体／
サンセリフ体、明朝体／ゴシック
体に分けられる。
- （おおまかな傾向として）
セリフの引っかけり → **本文向き**
サンセリフのシンプルさ
→ **サイン向き**

セリフ体・明朝体

a A あ ア 亜



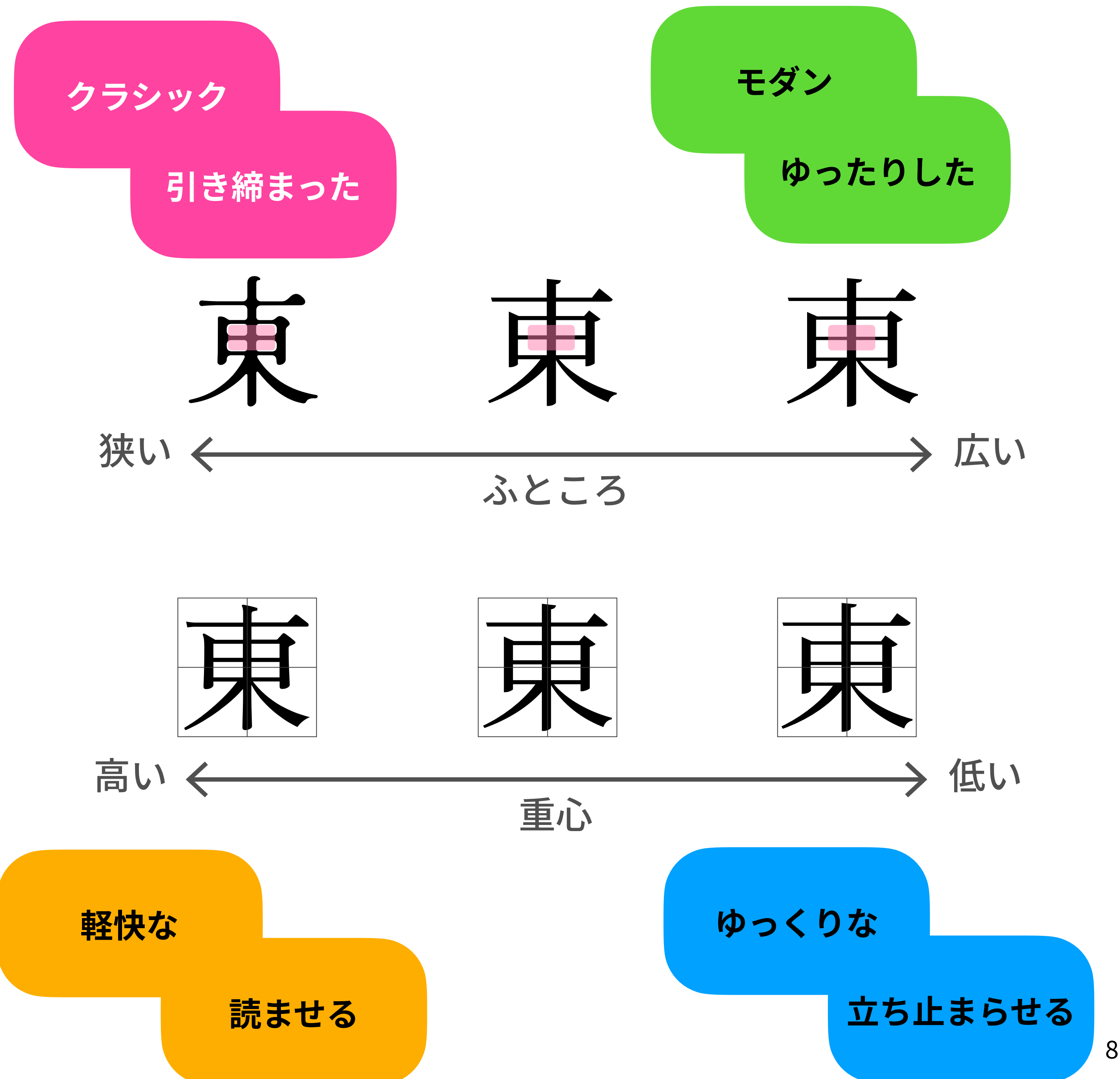
サンセリフ体・ゴシック体

a A あ ア 亜



ふところと重心

- ストロークで囲まれた空間
→ ふところ
- 文字の視覚的中心の高さ
→ 重心



ふところが狭い

游ゴシック

東京駅のプラットフォームに下り立つ

Noto Sans JP

東京駅のプラットフォームに下り立つ

新ゴ

東京駅のプラットフォームに下り立つ

ふところが広い

学参フォント、教科書体

- 活字はさまざまな理由で手書きの字体と異なる。
→ そのままでは漢字学習の参考にできない。
- 学参フォントは学習指導要領の代表的な字形に準拠しており、漢字学習の参考にできる。
- 教科書体は手書きの楷書に近い学参フォント。

A P-OTF リュウミン Pr6N

ふと北に進む

A P-OTF 新ゴ Pr6N

ふと北に進む

G-OTF 常改リュウミン ProN (学参明朝体)

ふと北に進む

G-OTF 常改新ゴ ProN (学参ゴシック体)

ふと北に進む

G-OTF 常改リュウミン ProN

心を込めて食事する人達

G-OTF 常改教科書 ICA ProN

心を込めて食事する人達

ユニバーサルデザイン（UD） フォント

誰にでも判別しやすい文字
を目指したフォント。

● ぼやけても間違いにくい字形

6 S 8 6 S 8

6 S 8 6 S 8

6 S 8 6 S 8

新ゴ

BIZ UD ゴシック

● 濁点・半濁点の衝突を回避

プ プ

Noto Sans JP

BIZ UD ゴシック

● ふところが広いシンプルな字形

あ あ

Noto Sans JP

BIZ UD ゴシック

UDを取り入れた教科書体と
して有名な**UDデジタル教科
書体（モリサワ）**

Webフォントと
美しく機能的な
タイポグラフィ



A $\xrightarrow{\text{符号化}}$ 1000001
 $\xleftarrow{\text{復号}}$

normal diagonal-fractions stacked-fractions Unicode (U+00BD)
1/2 ½ ½ ½

She might
She might

さまざまなフォント技術

ac



OpenType

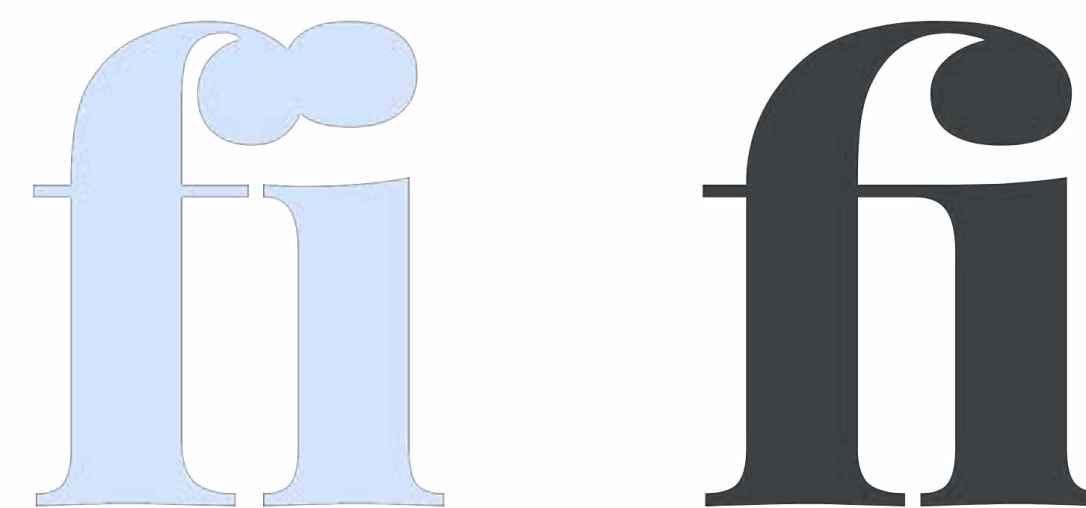
- マイクロソフト社とアドビ社により開発されたデジタルフォント規格。
- 合字、代替グリフ、プロポーショナルメトリクスなど、高度な組版のニーズを満たす機能を提供する。
- 特に**バリアブルフォント（Variable Fonts, VF, 可変フォント）**はより柔軟な組版を実現する次世代技術として注目されている。



合字／リガチャ

font-variant-alternates

- 隣り合う文字をつなげて新しい字形を与えたもの。
- 文字のエレメント同士が衝突することを防いだものや、装飾を目的としたものなどさまざま。
- OpenTypeではデフォルトで有効の常用合字のほかに、デフォルト無効の**任意的合字**や**歴史合字**などがある。



font-variant-ligatures: normal;

instinct

font-variant-ligatures: discretionary-ligatures;

instinct

使用フォント: Cormorant

数学記号の置換

合字のクリエイティブな使い方

プログラミング用途向けの等幅フォントFira Codeは、演算子を自然な数学記号に置き換えるのに合字機能を使っている。

font-variant-ligatures: none;

if (a != b || c >= d) i++;

font-variant-ligatures: normal;

if (a \neq b || c \geq d) i++;

手書きのゆらぎの表現

合字のクリエイティブな使い方

手書きフォントCaveatは文脈依存置換を用いて同じ文字でも文脈によってことなる字形を出し分ける。

→ 手書きの自然なゆらぎを再現している。

font-variant-ligatures: contextual;

Hello, everyone.

font-variant-ligatures: none;

Hello, everyone.

等幅フォントの文字幅調整

合字のクリエイティブな使い方

本来さまざまな幅を持つ文字を等しい幅に収めると、窮屈な文字や歪な空間が生まれてしまう。

→ GitHub発のMonaspaceは「余った空間を窮屈な文字に分けてあげる」という発想で、等幅フォントながらも読みやすい表示を実現している (Texture Healing)



代替グリフ

`font-variant-alternates`

- ひとつの文字に対して複数のグリフを定義し、切り替えできる。
- 一階建て／二階建て “a” のようにスタイルにまつわるもの (styleset) や、スワッシュ (swsh) などがある。

a a g g

使用フォント: IBM Plex Sans

Work Hard

Work Hard

使用フォント: Playball

東アジア言語字形

font-variant-east-asian

- 意味も読みも同じだが異なる字体を持つ漢字 → **異体字**
- 複数の字体を持つ文字がひとつのコードポイントにまとめられていることがある → **Han Unification**
- JISが定める字体の変遷によって変化したものは、JISのバージョンを指定することで切り替えができる。

font-variant-east-asian: normal;

葛

font-variant-east-asian: jis90;

葛

使用フォント: Noto Sans JP

→ 異体字については
後のスライドで詳述

数字の制御

font-variant-numeric

- OpenTypeは数字についても細かくグリフを制御する枠組みを提供している。
- オールドスタイル／ライニングの切替、分数の表記、スラッシュ付ゼロ、序数標識など

font-variant-numeric: oldstyle-nums;

O I 2 3 4 5 6 7 8 9

font-variant-numeric: lining-nums;

0 1 2 3 4 5 6 7 8 9

使用フォント: Cormorant Garamond

normal

diagonal-fractions

stacked-fractions

Unicode (U+00BD)

1/2 ½ ½ ½

使用フォント: Klee One

normal

slashed-zero

N0012 N0012

使用フォント: IBM Plex Sans

文字のスタイリング

- フォントにはボールド体（太字）やイタリック体（いわゆる斜体）が用意されているものがある。
- 後述するバリエアブルフォントが登場するまでは、スタイルとフォントファイルが一対一対応するものだった。

Roboto *Thin & Thin Italic*
Roboto *Light & Light Italic*
Roboto *Regular & Regular Italic*
Roboto *Medium & Medium Italic*
Roboto *Bold & Bold Italic*
Roboto *Black & Black Italic*

⚠ 対応するスタイルが用意されていないフォントに適用すると、擬似ボールド／イタリックが発生してグリフが潰れることがある。

Noto Sans JP 700

電撃解散は青天の霹靂

Noto Sans JP 400 (Faux Bold)

電撃解散は青天の霹靂

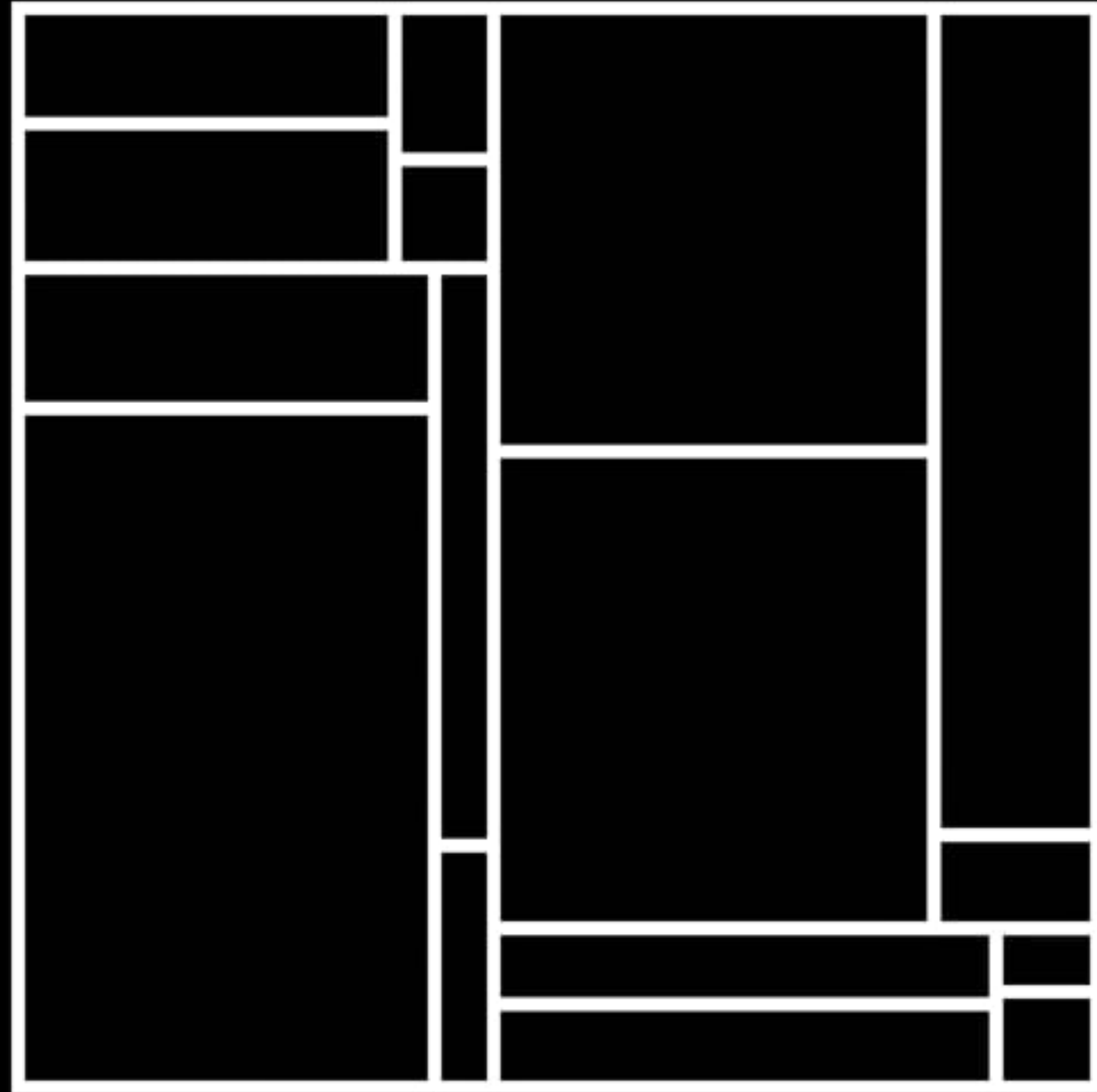
バリエブルフォント

- 複数のスタイルをひとつのフォントファイルに収める技術
- **バリエーション軸**: 調整可能なスタイルの値の許容範囲を定義したもの
- **例: 太さ軸 (wght) → 100-900**
400が標準、700がボールドに対応する。



GoogleSansFlex_9pt-Black.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_9pt-Bold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_9pt-ExtraBold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_9pt-ExtraLight.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_9pt-Light.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_9pt-Medium.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_9pt-Regular.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_9pt-SemiBold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_9pt-Thin.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_24pt-Black.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_24pt-Bold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_24pt-ExtraBold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_24pt-ExtraLight.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_24pt-Light.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_24pt-Medium.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_24pt-Regular.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_24pt-SemiBold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_24pt-Thin.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_36pt-Black.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_36pt-Bold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_36pt-ExtraBold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_36pt-ExtraLight.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_36pt-Light.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_36pt-Medium.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_36pt-Regular.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_36pt-SemiBold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_36pt-Thin.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_72pt-Black.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_72pt-Bold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_72pt-ExtraBold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_72pt-ExtraLight.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_72pt-Light.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_72pt-Medium.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_72pt-Regular.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_72pt-SemiBold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_72pt-Thin.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_120pt-Black.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_120pt-Bold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_120pt-ExtraBold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_120pt-ExtraLight.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_120pt-Light.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_120pt-Medium.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_120pt-Regular.ttf	2025年12月12日 7:49	127 KB	TrueTy...フォント
GoogleSansFlex_120pt-SemiBold.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex_120pt-Thin.ttf	2025年12月12日 7:49	128 KB	TrueTy...フォント
GoogleSansFlex-Vari...sz,slnt,wght.ttf 2025年12月12日 7:49 4 MB			

54 → 1ファイル
8.2 → 4MB





TIME: 1



TIME: 20



TIME: 40



TIME: 60



TIME: 80



TIME: 100

使用フォント: Anicons

CLIMATE

1979

CLIMATE

1999

CLIMATE

2000

CLIMATE

2010

CLIMATE

2019

CLIMATE

2030

CLIMATE

2040

CLIMATE

2050

使用フォント: Climate Crisis

オプティカルサイズ

opsz軸

文字サイズに合わせて字形を調整する技術。

CSSではfont-optical-sizingとして利用でき、デフォルト値は有効(auto)。

バリエブルフォントとともに再び一般的に用いられるように。

Their distinct,
feverish sound

An astute and diverse release

But if not conscience stricken, Sabinsport
was interested in the discussion. It stirred
deeper than Ralph in his disgust with the
situation had dreamed. Letters to his Pro
Bono Publico column flowed in daily.

© All rights reserved.

text

text

text

text

a a a a a a a a



A $\xrightarrow{\text{付与化}}$ 1000001
 $\xleftarrow{\text{復号}}$

normal diagonal-fractions stacked-fractions Unicode (U+00BD)
1/2 ½ ½ ½

She might
She might

Google Fontsでできること

ac



TIME: 1



TIME: 20



TIME: 40



TIME: 60



TIME: 80



TIME: 100



Google Fonts

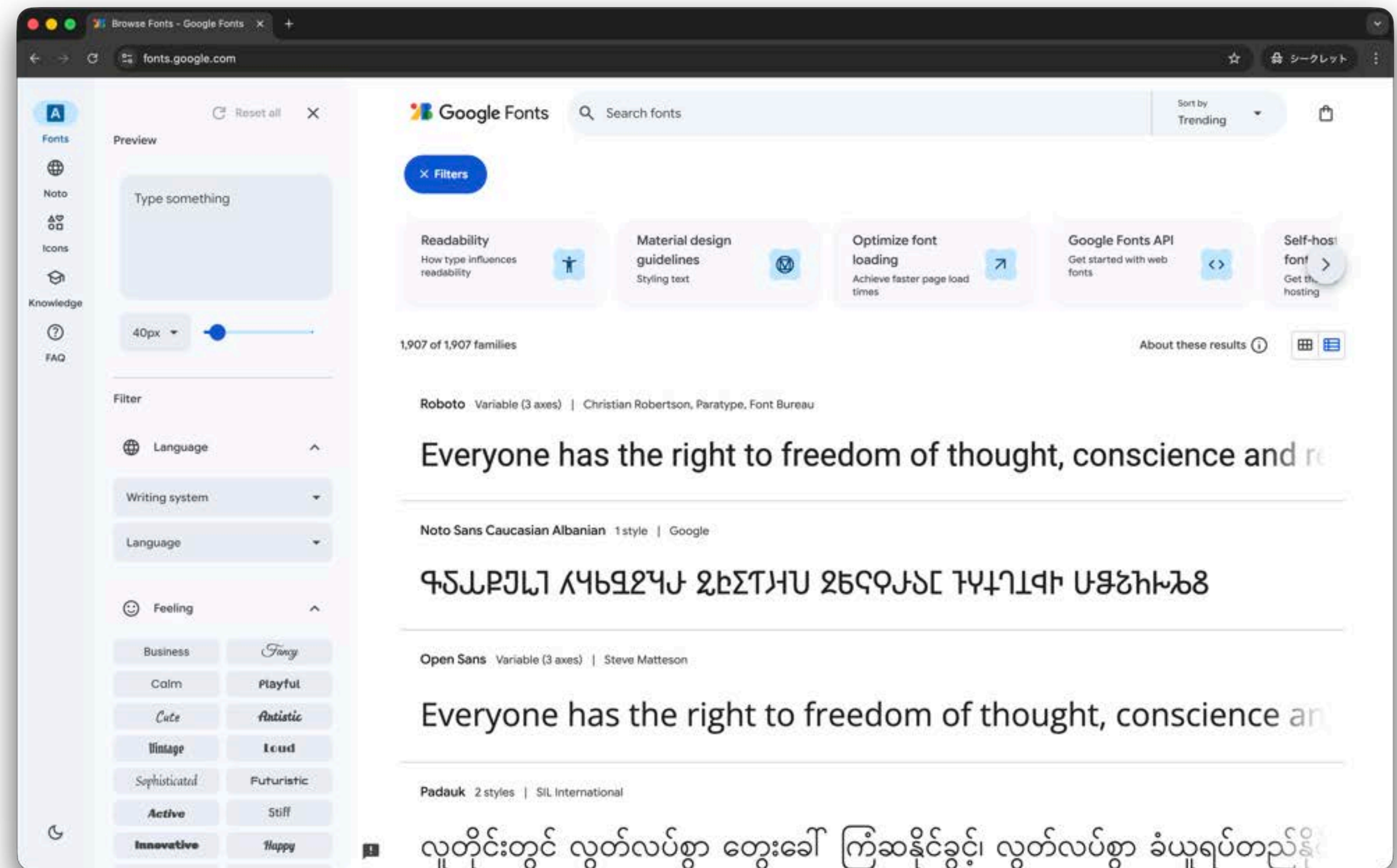
オープンソースフォントのプラットフォーム

世界中のオープンソースフォントを掲載したウェブサイト。

学習コンテンツやNotoプロジェクトの概要も提供する。

コンテンツはGitHubで管理されており、誰でも貢献可能。

<https://github.com/google/fonts>



Google Sans Flex

Designed by [Google](#)[Learn more about variable fonts](#)

Select preview text:

Latin

Language

Variable Axes



Weight



Grade



Slant



Width



Custom



40px



Web font is awesome



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

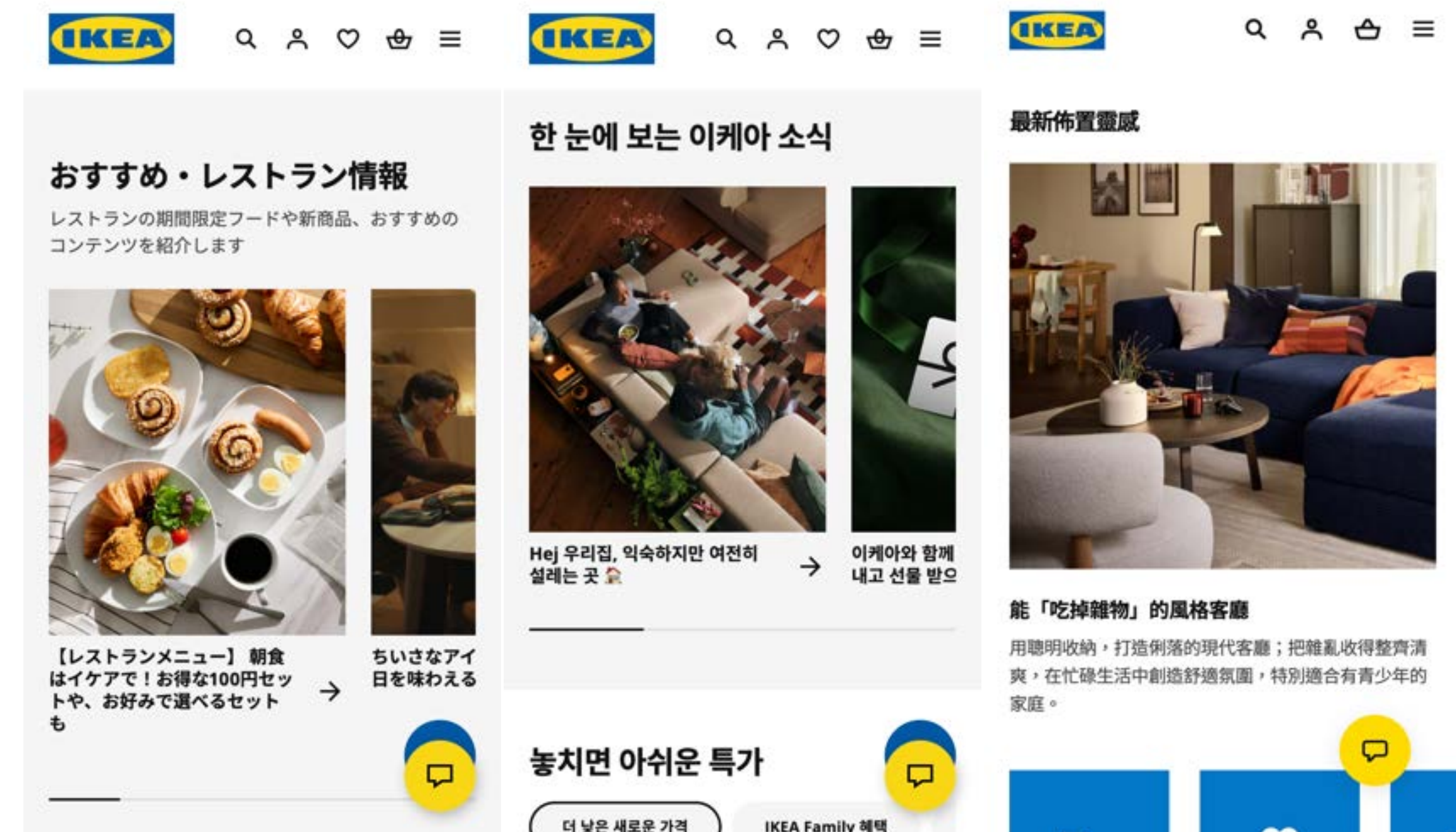
Styles used above

Size: 40 Optical Size: Auto Slant: 0 Width: 100.4 Weight: 401

Size: 22 Optical Size: Auto Slant: 0 Width: 100 Weight: 400 Heading 2

Notoプロジェクト

- 世界中の文字を表示することを目的としたプロジェクト
- No Tofu: 豆腐文字をなくす
- 現時点で1000を超える言語、150を超える文字体系に対応
- ひとつの書体ファミリーとしてデザインを共有しているため、一貫性のある国際化デザインが可能になる



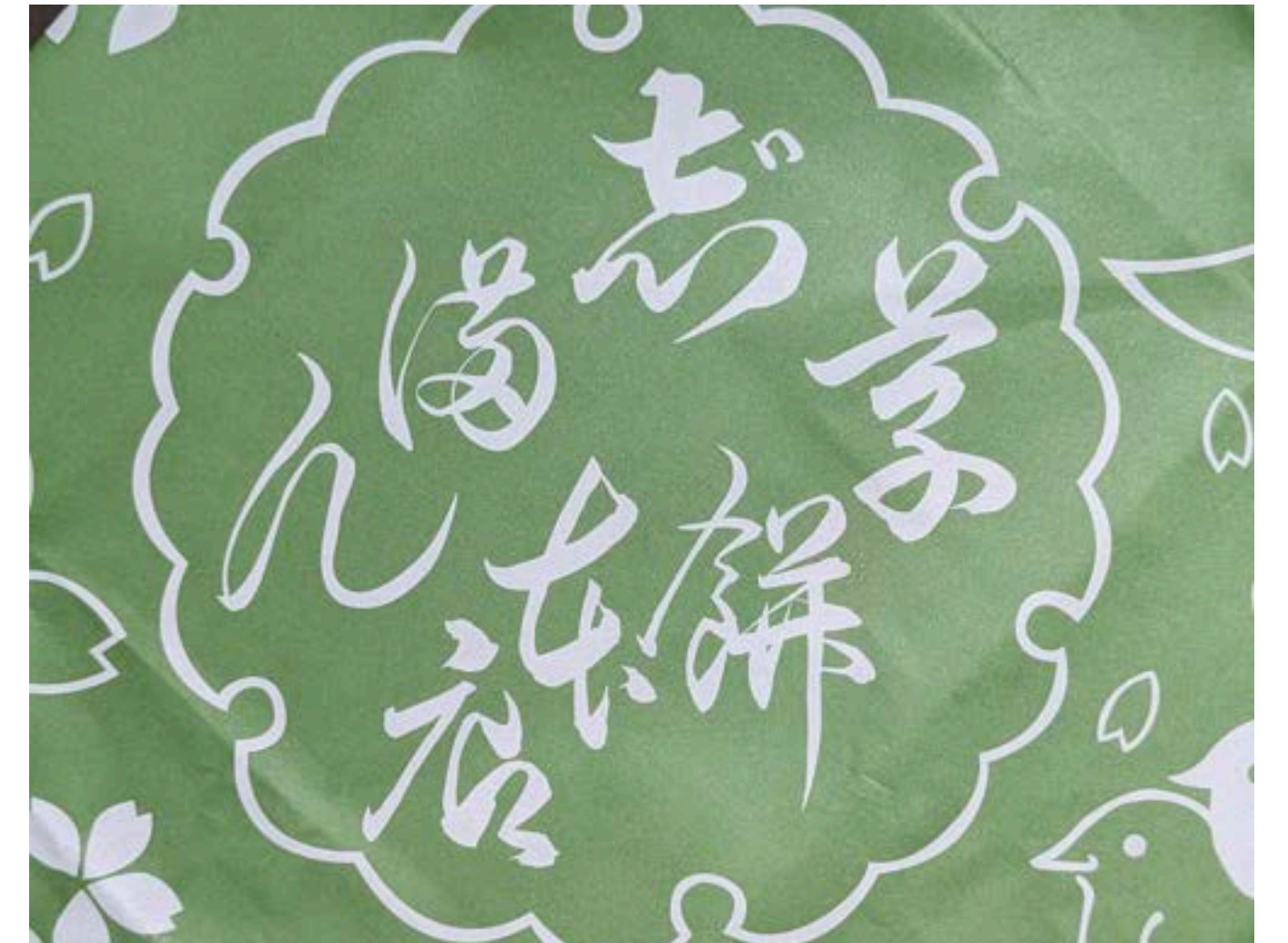
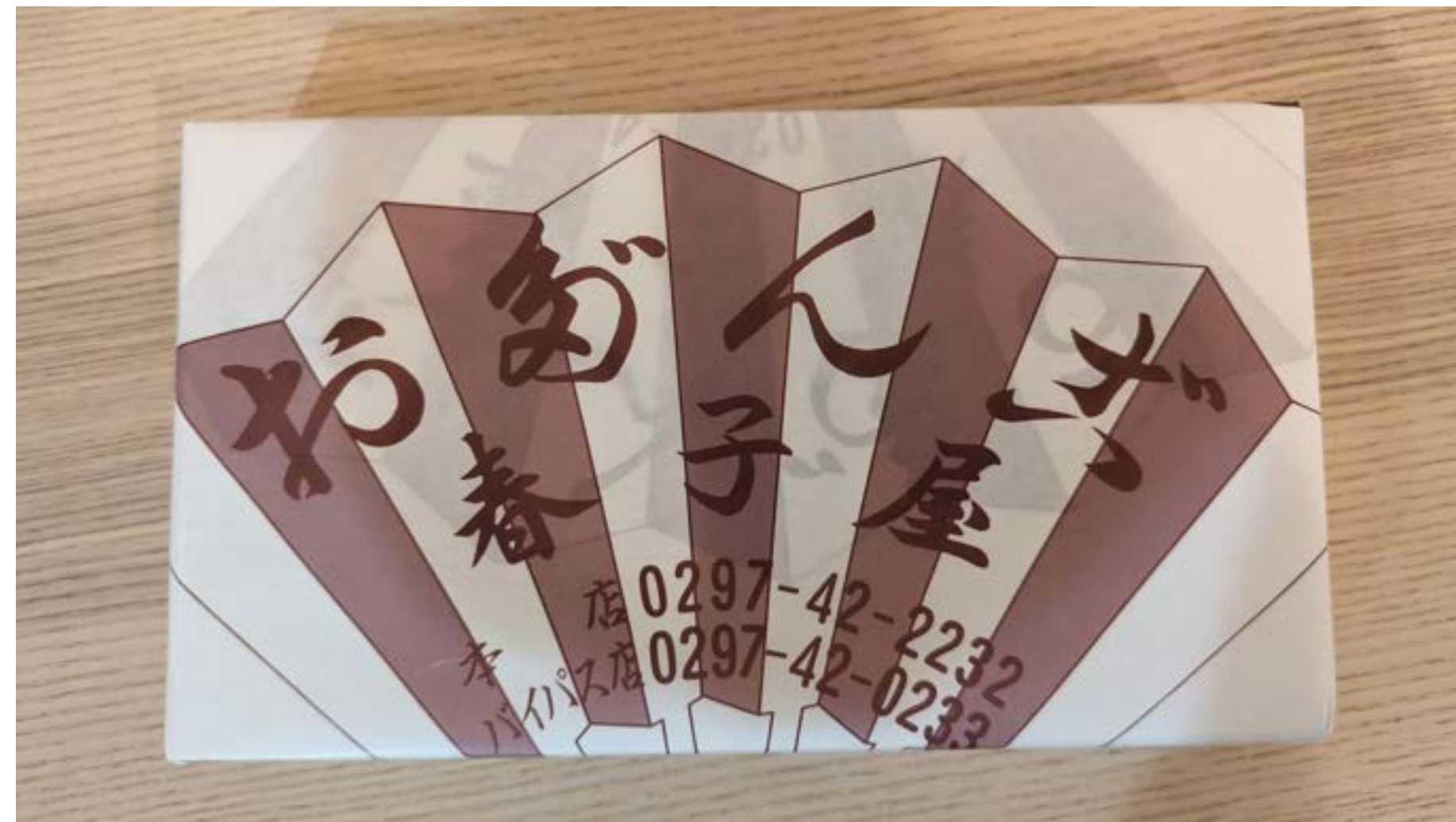
日本、韓国、台湾のIKEAウェブサイトのスクリーンショット

Noto Serif Hentaigana

- Unicodeに登録された変体仮名
286文字をサポートするセリフ体
フォント。
- 7種類のウェイトに対応。

以移波に保倍亦 予初勢冻文

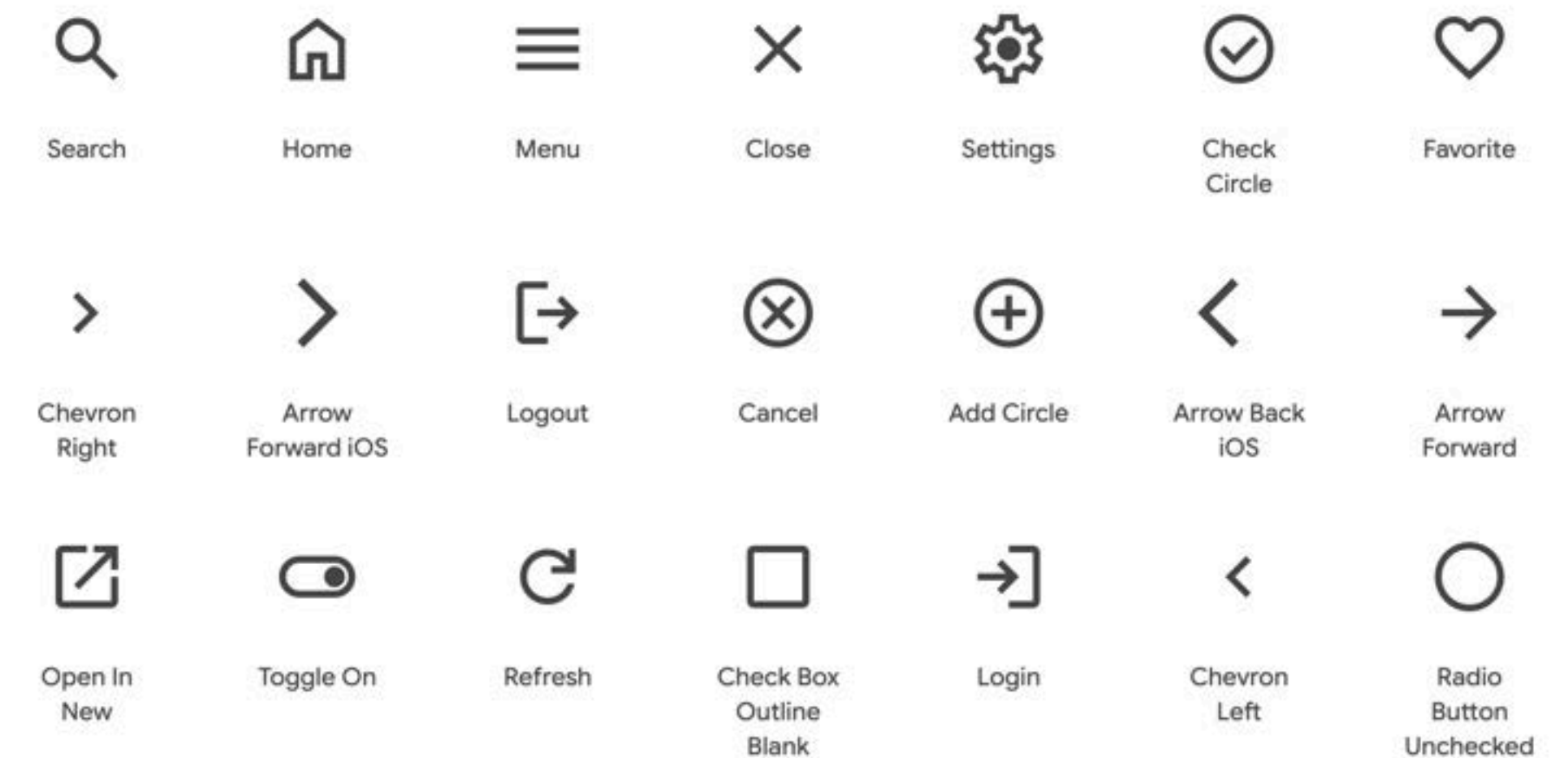
以移波に保倍亦 予初勢冻文 初徳代毎匙取 以迄名らむ 以移
予初勢冻文 初徳代毎匙取 以迄名らむ以移波に保倍亦 予初勢
代毎匙取 以迄名らむ 以移波に保倍亦 予初勢冻文 初徳代毎



Material Symbols

アイコンフォントとその応用

- 文字だけでなく、アイコンを表示するためのフォントも提供している。
- Material Symbols: デザインシステム
Material Designの一部として開発されたアイコンセット。
- ウェブに限らず、印刷物やサインなどさまざまなシーンで活用されている。



ニューヨーク近代美術館（MoMA）の館内表示

Webフォントの基本的な使い方

HTML / CSS スニペットのサンプル

- Linkタグで読み込みたいフォントを指定する。
- フォントを適用したい要素のCSSにfont-familyを指定する。
- 他にも効率的にフォントを読み込むためのさまざまなパラメーターが用意されている。

```
<link href="https://fonts.googleapis.com/css2?family=Noto+Sans+JP:wght@100..900&display=swap" rel="stylesheet">
```

```
<style>
.example {
  font-family: 'Noto Sans JP', sans-serif;
  font-weight: 400;
}
</style>
<span class="example">
  こんにちは
</span>
```

こんにちは

アイコンフォントの使い方

HTML / CSS スニペットのサンプル

- 基本的な使い方は同様。
- “search” という文字列を検索のアイコン（グリフ）に紐づけるところで合字の技術が使われている。
- バリアブルフォントとして太さや塗りつぶしの制御が可能。

```
<link rel="stylesheet" href="https://  
fonts.googleapis.com/css2?  
family=Material+Symbols+Outlined:opsz,wght,  
FILL,GRAD@20..48,100..700,0..1,-50..200" />
```

```
<style>  
.example {  
  font-family: 'Material Symbols Outlined';  
  font-variation-settings:  
    'FILL' 0,  
    'wght' 400,  
    'GRAD' 0,  
    'opsz' 24  
}  
</style>  
<span class="example">search</span>
```





A $\xrightarrow{\text{符号化}}$ 1000001
 $\xleftarrow{\text{復号}}$

normal diagonal-fractions stacked-fractions Unicode (U+00BD)
1/2 ½ ½ ½

She might
She might

異体字の解剖

ac

≡ TIME: 1
≡ TIME: 20
∧ TIME: 40
× TIME: 60
+ TIME: 80
× TIME: 100



身の回りの異体字

- 人名や地名には読みと意味が同じだが字体が違う文字、**異体字**が溢れている。
- 異体字の制御には複数の要因が絡まっていることがあり、ややこしい。

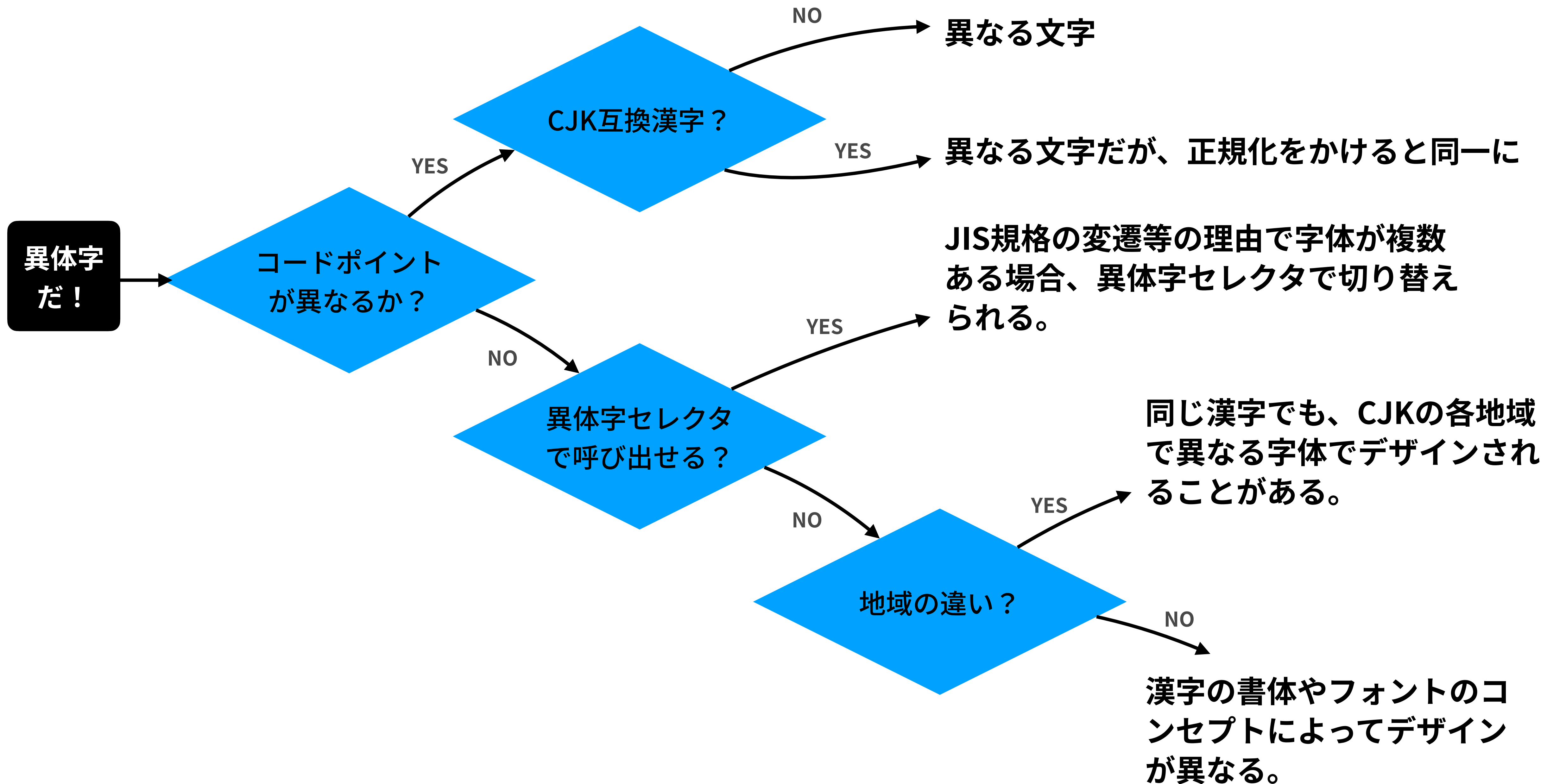
高橋 高橋

吉田 吉田

宝塚 宝塚

葛飾 葛城

令和 令和



コードポイントとCJK互換漢字

- 異なるコードポイントが割り当てられているケースは多い。
 - 高 (U+9AD8) と高 (U+9AD9)
 - 吉 (U+5409) と吉 (U+20BB7)
 - 塚 (U+585A) と塚 (U+FA10)
- CJK互換文字は等価とみなされる漢字。
 - たとえば「塚」を検索にかけると「塚」と「塚」の両方がヒットする。
→ Unicode正規化

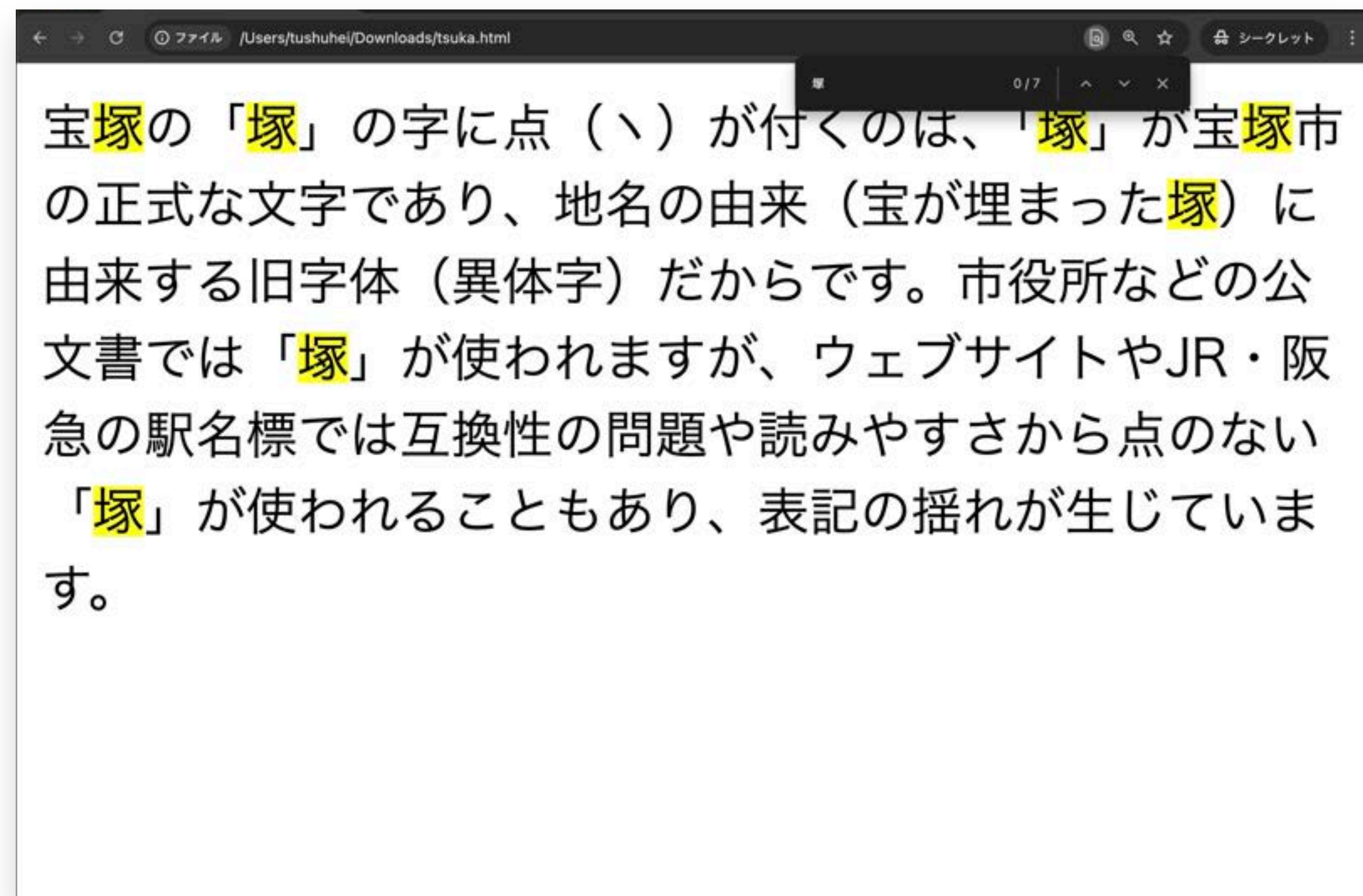
U+30EDE

邇

U+30EDE

Tab キーで選択

Google日本語入力では“U+”にコードポイントを続けることで、その文字を入力できる。



「塚」と「塚」の両方がヒットする様子。

JISの変遷と例示字形

- 漢字の字形はそのフォントが制作された当時のJISが定める例示字形に影響を受けている。
- モリサワのNシリーズ→JIS2004に対応
- 複数のJIS規格の例示字形を持つフォントもあり、CSSではfont-variant-east-asianプロパティで呼び出せる。

葛 葛 葛

Osaka ヒラギノ角ゴ Pro ヒラギノ角ゴ ProN

Code point	jis78	jis83	jis90	jis04
U+845B	葛	葛	葛	葛
U+698A	榊	榊	榊	榊
U+9771	靱	靱	靱	靱
U+53A9	厩	厩	厩	厩
U+990C	餌	餌	餌	餌

異体字セレクタ

Variation Selector, VS

- 直前の文字の字形を変えるために使用される特殊な文字。基底文字につなげて**異体字シーケンス**を構成する。
- 漢字異体字シーケンス（IVS）で利用できる異体字は異体字データベース（IVD）に収められている。
- たとえば
葛 (U+845B) + **VS17** (U+E0100) = **葛**

葛

E0100
Adobe-Japan1
CID+1481

葛

E0101
Adobe-Japan1
CID+7652

葛

E0102
Moji_Joho
MJ022335

葛

E0103
Moji_Joho
MJ022336

葛

E0104
Moji_Joho
MJ022340

葛

E0105
Moji_Joho
MJ022341

葛

E0106
Moji_Joho
MJ022338

葛

E0107
Moji_Joho
MJ022337

葛

E0108
Moji_Joho
MJ022339

IVDの「葛」の項目

地域による字体の違い

- **CJK**: Chinese, Japanese, Korean
→ 漢字およびその派生文字の書記体系
- 同じコードポイントを持つ漢字でも、地域によって用いられる字体が異なる。
- Noto CJKのようなPan-CJKフォントはすべての字体を含む。
→ **利用時の適切な言語設定が肝要**

Unicode	日本語	中国語 (簡体字)	中国語 (繁体字、台湾)	中国語 (繁体字、香港)	韓国
U+4EE4	令	令	令	令	令
U+9AA8	骨	骨	骨	骨	骨
U+66DC	曜	曜	曜	曜	曜
U+6B21	次	次	次	次	次

Noto Sans CJKの地域別の字形の違い

縦置き充電器
縦置き充電器

日本語字形と中国語簡体字字形の違い

書体のコンセプトによる違い

- フォントが元にする書体によって字体が変わることもある。
- 印刷効率化のために変化した活字由来の字体と、手書き字体の違い
- 学参フォントでも「令」の字体になる。



Noto Serif JP



Klee One



Zen Kurenaido



A $\xrightarrow{\text{符号化}}$ 1000001
 $\xleftarrow{\text{復号}}$

normal diagonal-fractions stacked-fractions Unicode (U+00BD)
1/2 ½ ½ ½

She might
She might

BudouX

機械学習とタイポグラフィ

ac

≡ ≡ ≡ ≡ ≡ ≡
TIME: 1 TIME: 20 TIME: 40 TIME: 60 TIME: 80 TIME: 100



日本語における自然な改行を目指して

分かち書きをしない言語が持つ課題

- 日本語では単語や文節の途中で改行が挿入されることがある。
→ **洗練さに欠ける、読みにくい**
- 画面幅が流動的なWebでは、改行位置を指定できない。
→ **改行可能位置の指定**
- 読みやすい文章は、文節の間で改行することが多い。
→ **文節の分かち書きが必要**



単語の途中で改行が挿入される例

BudouX: 軽量な文節分かち書きエンジン

<https://github.com/google/budoux>

- オープンソースプロジェクト
- 機械学習モデル含めて15KBと軽量

“今日は牛
丼を食べ
た。”

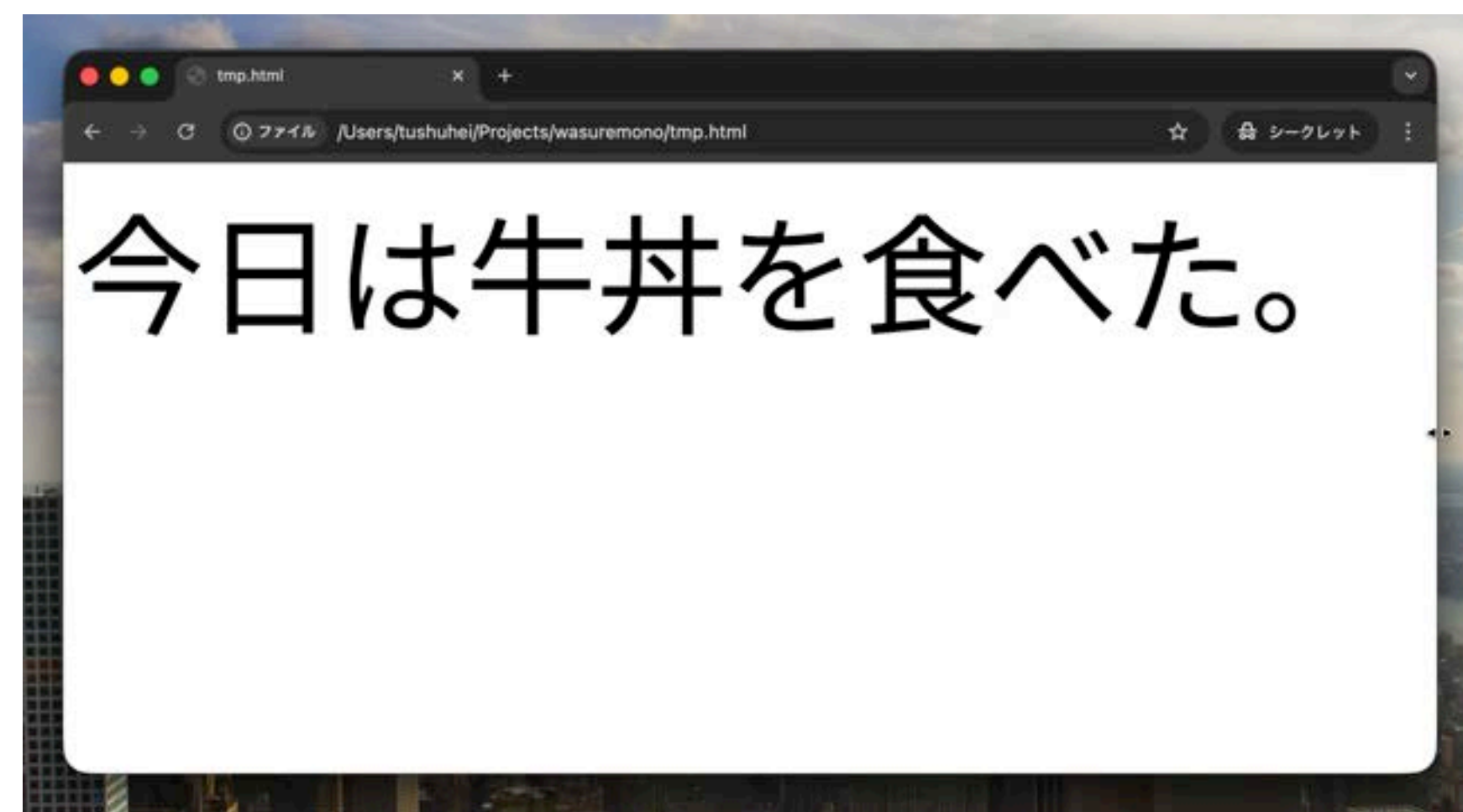
- 今日は
- 牛丼を
- 食べた。

文節分かち書き

HTMLマークアップ

レンダリング

```
<style>
  .example {
    word-break: keep-all;
    overflow-wrap: anywhere;
  }
</style>
<span class="example">
今日は&ZeroWidthSpace;牛丼を
&ZeroWidthSpace;食べた。
</span>
```



問題の定式化

- 与えられた文章の各文字に対して、文末の先頭か否かを推定する問題として捉える。
→ 二値分類問題

入力 **今日は牛丼を食べた。**

機械学習システム

出力 1 0 0 1 0 0 1 0 0 0

特徴量エンジニアリング

- 「食」が文節の頭であると推定するには、その前後情報が重要。
→ 文脈を入力変数（特徴量）に落とし込む必要がある。

入力 今日は牛丼を**食**べた。

機械学習システム

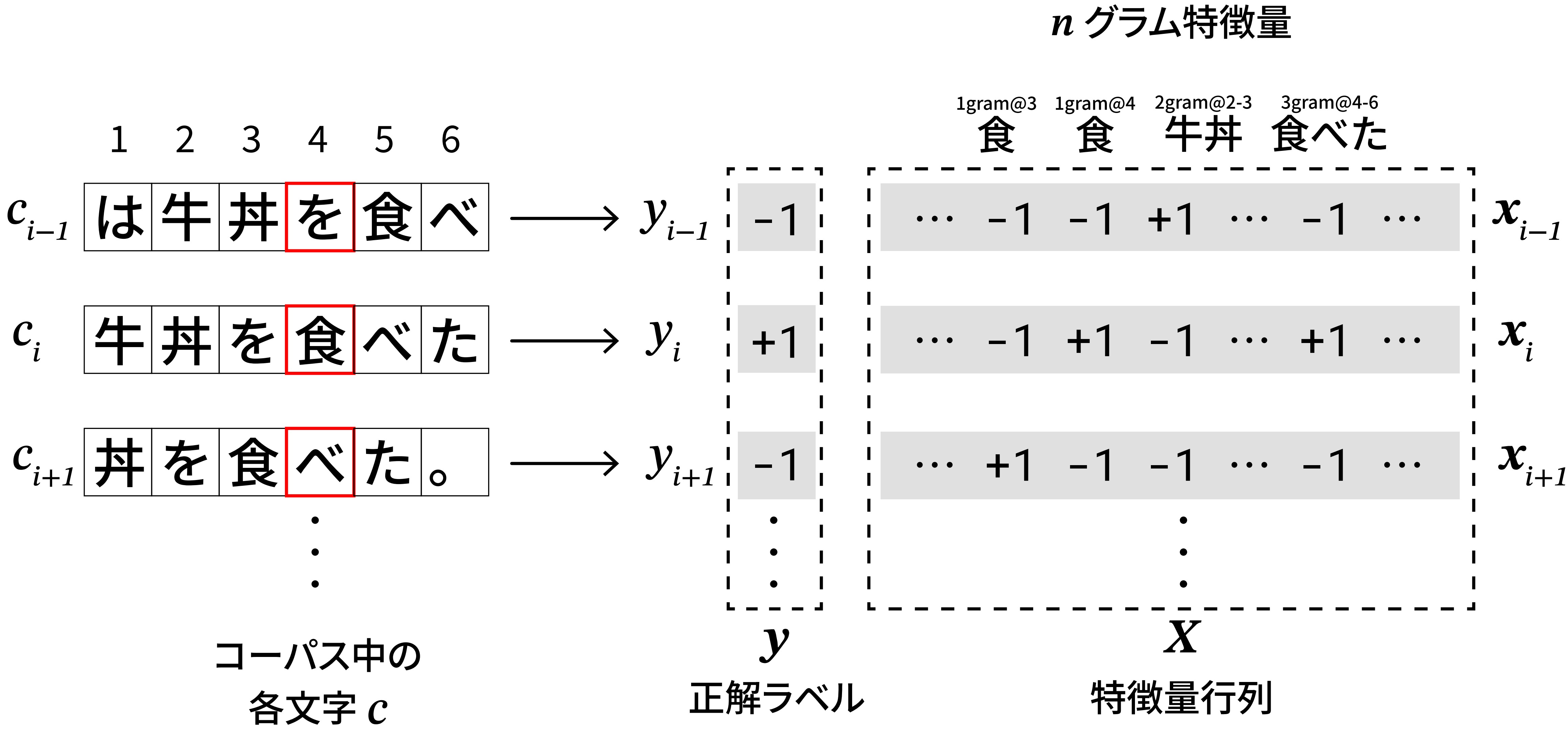
出力 1 0 0 1 0 0 1 0 0 0

[illegible]

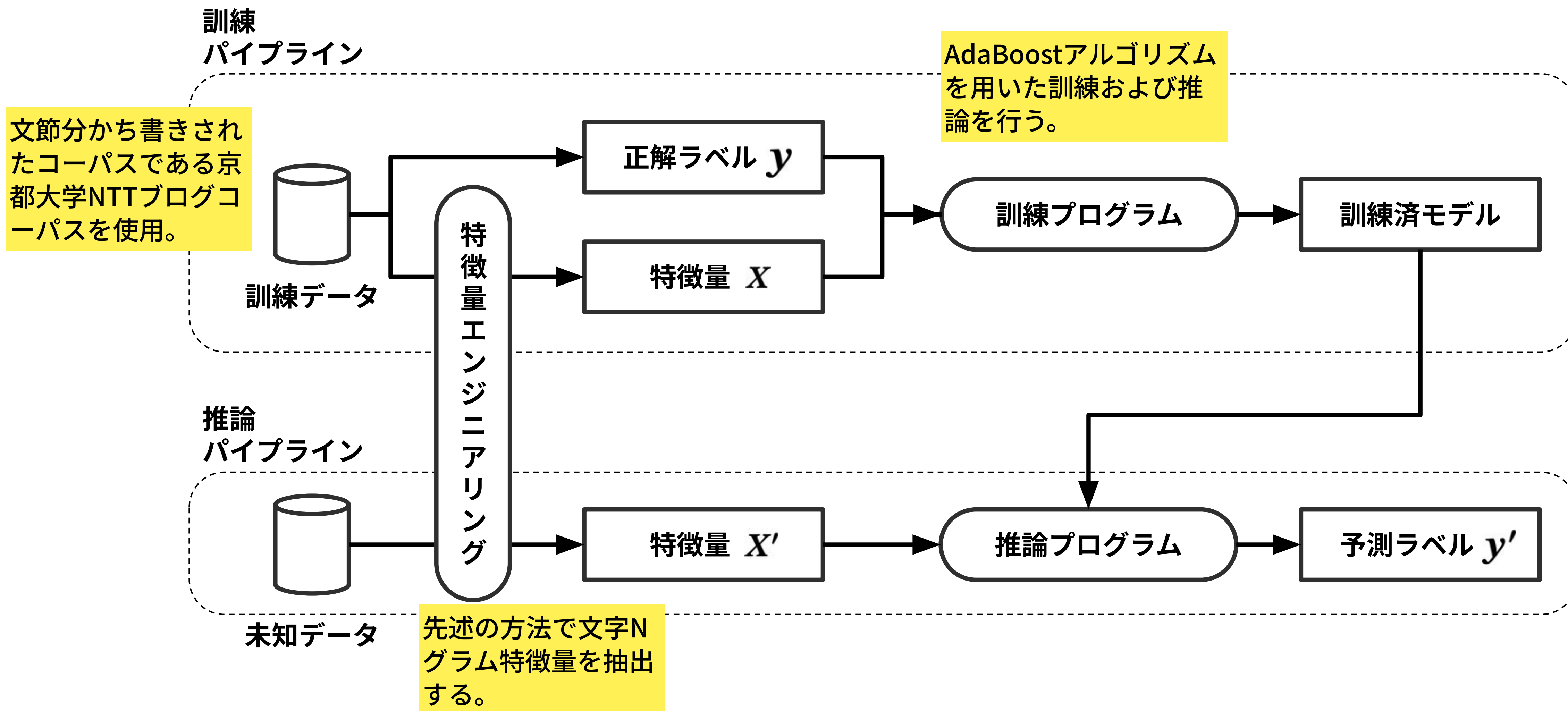
機械学習システム

出力

データ加工の全体像



機械学習のプロセス



機械学習モデルの概観

- BudouXで出力したAdaBoost機械学習モデルは、文字Nグラムと対応する貢献度スコアのペアで表される。
- 対象の文字の周辺に現れる文字Nグラムについて、貢献度スコアの和が正なら文末先頭、負なら文末先頭でない（＝前の文字と接続する）。
- JSONフォーマット
 - 第1キー：文字Nグラムの種類
 - 第2キー：文字Nグラムの値
 - 値：貢献度スコア

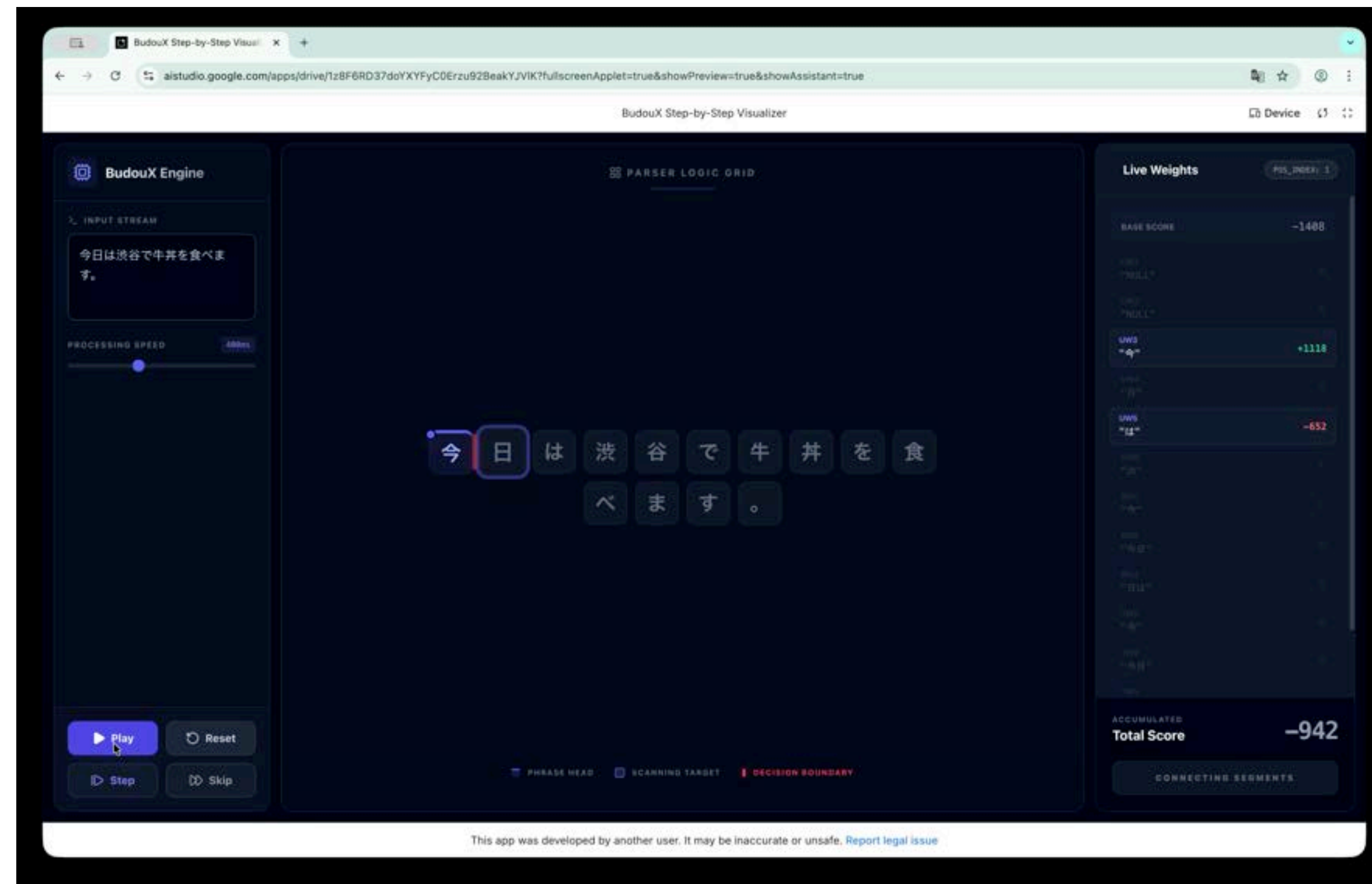
```
{
  "UW3": {
    "、": 4698,
    "。": 6235,
    "の": 3583,
    "に": 3682,
    "は": 3987,
    "を": 5216,
    "が": 3860,
    ...
  },
  ...
  "TW3": {
    "という": 737,
    "ている": -1516,
    "ところ": -944,
    ...
  }
}
```

機械学習ランタイム

- 機械学習ライブラリ等を必要とせず、マッチする文字Nグラムに対応する貢献度スコアの和を取るのみで実装できる。

```
function parse(str: string, model: Map<string, Map<string, number>> ) {
  const result = [str[0]];
  for (let i = 1; i < str.length; i++) {
    let score = -0.5 * [...model.values()]
      .map(group => [...group.values()])
      .flat()
      .reduce((prev, curr) => prev + curr, 0);
    score += model.get('UW1')?.get(str.substring(i - 3, i - 2)) || 0;
    score += model.get('UW2')?.get(str.substring(i - 2, i - 1)) || 0;
    score += model.get('UW3')?.get(str.substring(i - 1, i)) || 0;
    score += model.get('UW4')?.get(str.substring(i, i + 1)) || 0;
    score += model.get('UW5')?.get(str.substring(i + 1, i + 2)) || 0;
    score += model.get('UW6')?.get(str.substring(i + 2, i + 3)) || 0;
    score += model.get('BW1')?.get(str.substring(i - 2, i)) || 0;
    score += model.get('BW2')?.get(str.substring(i - 1, i + 1)) || 0;
    score += model.get('BW3')?.get(str.substring(i, i + 2)) || 0;
    score += model.get('TW1')?.get(str.substring(i - 3, i)) || 0;
    score += model.get('TW2')?.get(str.substring(i - 2, i + 1)) || 0;
    score += model.get('TW3')?.get(str.substring(i - 1, i + 2)) || 0;
    score += model.get('TW4')?.get(str.substring(i, i + 3)) || 0;
    if (score > 0) {
      result.push(str[i]);
    } else {
      result[result.length - 1] += str[i];
    }
  }
  return result;
}
```


仕組みの可視化



<https://bit.ly/budoux-vis>

さまざまなプログラミング環境への展開

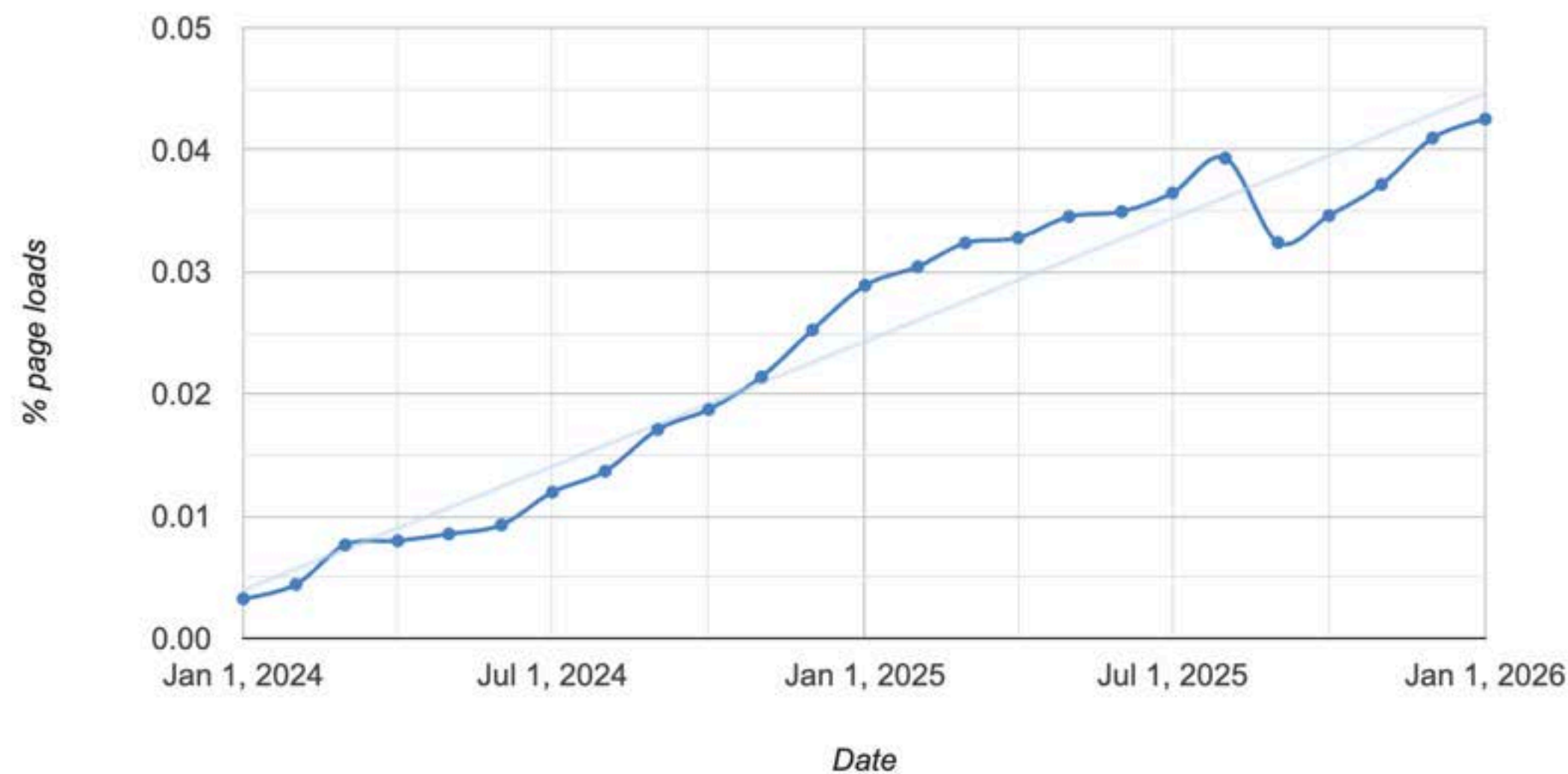
- 公式レポジトリではPython, Java, TypeScript (JavaScript) の3つのポートを提供。
- GitHubコミュニティによって、Go, Rust, Dartなどさまざまな言語に移植されている。
- ICU (International Components for Unicode) にも
MlBreakEngineとして搭載されている
→ **AndroidやChromeなど幅広いプラットフォーム**
で利用可能。

word-break: auto-phrase

- Chrome 119以降では `word-break: auto-phrase` がサポートされ、ja 言語設定化で日本語のフレーズ単位での改行が可能に。

```
<html lang="ja">
<style>
h1 {
  word-break: auto-phrase;
}
</style>

<h1>窓ぎわのトットちゃん</h1>
```



Chromeのページロードにおいて
`word-break: auto-phrase`が占める割合の推移

`word-break: normal`

窓ぎわのトッ
トちゃん

`word-break: auto-phrase`

窓ぎわの
トットちゃん

さまざまな言語のサポート

- BudouXのコンセプト: Simple, Standalone, **Language-neutral**
- 文字N-gram特徴量のみを利用し、特定の言語に特有の要素を学習に用いないようにしている。
- 現時点で日本語の他に中国語とタイ語をサポート。
 - ただし、日本語以外の言語は文節（フレーズ）ではなく単語区切り。

まとめ

- 文字とフォントの基礎から、いくつかの最新のWebタイポグラフィ技術まで概観した。
- 毎日目にする基本的な要素だが、それを支える技術は日々進化している。
- 機械学習／AIによって、美しいタイポグラフィをスケール可能な方法で提供していきたい。

