# Multilingual Document Recognition
# Erkennung Multilingualer Dokumente
# Reconnaissance de Documents Polyglotes


## Larry Spitz

*Daimler Benz Research and Technology Center*
*1510 Page Mill Road*
*Palo Alto, CA 94304 USA*
spitz@rtna.daimlerbenz.com


254 STYLISTIQUE COMPARÉE

message : "Don't stoop! : Tenez-vous droit!" Reste à défendre le changement de point de vue ; il est certain que "stoop" est plus direct que "se pencher", et que l'ordre : "Ne vous penchez pas" serait ambigu ; en fait, "stoop" exprime une attitude, qu'il faudrait rendre par une phrase négative : "ne vous tenez pas penché comme ça, ne vous voûtez pas ainsi". Nous serions ici, comme c'est fréquemment le cas, dans un domaine à cheval sur la structure (difficulté d'utilisation de "se pencher") et sur la métalinguistique : préférence pour les ordres positifs : "Taisez-vous !" plutôt que : "Ne parlez pas!", donc "Tenez-vous droit" plutôt que "Ne soyez pas penché".

Dans la presque totalité des exemples, l'explication de la modulation est à chercher exclusivement dans la métalinguistique. Soit le verbe "to fill" : le français n'emploie guère cette image pour indiquer une satisfaction gastronomique : "Je suis plein, je suis rempli" ne se disent guère ; "j'ai le ventre plein" est vulgaire. Il faut voir là une conséquence de la "civilité puérile mais honnête" de nos grands-parents, auxquels répugnaient certains mots tels que "manger", "boire", "digérer", etc. Ces tabous culturels obligent le traducteur à des changements de point de vue qui n'ont aucune excuse structurale, et qui par conséquent ne s'imposent pas à tous les usagers de LA avec une force égale. L'affiche de Coca-Cola qui proclame : "Coca-Cola refreshes without filling" (variante: "Coca-Cola does not fill") ne peut se traduire littéralement, d'autant plus que le sens de "fill" est subjectif. Les traducteurs canadiens ont bien vu la nécessité d'une modulation : "La boisson légère, qui rafraîchit." C'est une modulation par renversement, dont on peut trouver de nombreux exemples.

中國古時候，有一個人姓馬，名字叫書田。他這個名字是他父親給他起的。爲甚麼給他起了這個名字呢？那是有原因的。

他父親從前是做官的，做過不大不小的官，一共做了十幾年。在那個時候做官的，如果做了很多年，一定是很有錢。他父親本來也應該有一點兒錢，可是到了他父親不做官回到老家的時候，手裏是沒有多少錢了。他爲甚麼沒有多少錢了呢？一本是因爲他父親是一個不要錢的官，從來不跟老百姓要錢。要是做官的不要老百姓的錢，到了不做官的時候，當然不會有很多的錢了。二本是因爲他父親有一點兒老毛病，這個老毛病也可以說是好的毛病，也可以說是不好的毛病。是甚麼毛病呢？就是他父親手裏要是有了錢，就要用。他做甚麼用呢？他喜歡買東西。別的東西不買，他就喜歡買書。在他父親做官的十幾年裏，有了錢就買書，所以他沒有多少錢，可是，買的書實在不少。

他父親不做官回到老家以後，看看家裏有這麼多的書，心裏很喜歡。他父親對他母親說：“我實在喜歡書。書，實在是好東西。現在，你看，我們家裏有這麼多的書，你心裏一定也該歡喜了。”他母親看了看他父親，然後說：“我看你怎麼喜歡，我當然也很喜歡。不過，別人做官的，到了不做官的時候，回到老家去，多半都是買了很多田地，以後可以過日子。現在，我們只有這麼多的書，能當飯吃嗎？”他父親說：“我們雖然沒有田，可是，有這麼多的書，要是念了書，那能沒飯吃呢？從前古人說過：「書中自有千鐘粟。」「有千鐘粟」就是有飯吃的意思。所以我可以說，這書就是我們的田。”

過了幾天，他母親生了一個兒子，就是馬書田。因爲從前他母親生的都是女兒，這是頭一次生兒子，他父親、母親兩個人都喜歡的不得了。他母親就叫他父親給他這個兒子起一個名字。這時候，他父親想了一想，又往前後看了一看。他看見了這麼多的書，又想前幾天說過的話：“這書就是我們的田地。”他父親就對他母親說：“我想出名字來了，我現在寫給你看看。”他父親就先寫一個「書」字。他母親看是一個「書」字，心裏想：“這個人別的不想，他的心裏就是想書。可是，我從來不喜歡的就是書。”他父親才要說出來他的意見的時候，看見他父親又往下寫了一個字。他就在「書」字的下頭又寫了一個「田」字。他母親看見了太喜歡了，說：“田是我最希望有德，我希望兒子的名字有「書」字也有「田」字。希望我兒子將來有書也有田地，不要像你，就有書沒有田地。”從這個時候以後，他們的兒子就叫馬書田了。

馬書田很小的時候，就喜歡書。那時候，他雖然看不懂，可是常常的要看書。他父親看見他的兒子喜歡書，心裏有說不出來的歡喜。有一天，他父親說：“書田長大了以後，一定是一個好學生，將來也一定是一個很有學問的人。”

새로움을 느낀다 성활력소

새계에서 가장 정조관념이 굳다는 것이 중국의 여성들이다.

흔들리는 것도 적고 확실히 일본의 여성들보다는 정조관념이 강하다.

"메인 파즈"(할 수 없다)란 말의 표현과 같이 그녀들은 자기의 운명에 순종함은 물론, 제2, 제3의 부인으로도 좋다고하며 여자들 사이에 싸움도없다.

그렇다고 해서 그녀들은 성적으로 미숙하지도 않다. 그뿐아니라 그 신비에 찬성문제는 세계 제1위일 것이다. 그러나 그엇은 "말들어진 여성"이 란점에 중국 여성들의 특징이 있다. 중공정권수립 후 성의 개방, 자유가 되살아나서 여성 자신의 입김이 거세지고 있기는 하지만 아직까지 은밀한 곳에 서는 옛날과 같은 신비가 담겨져 있다.

"무엇인가 하나만은 남기고 벗는다"란 것도 사실은 최대의 효과를 염두에 둔 에로티즘이다.

그 원천은 유명한 "조인 발"에 있다고 하겠다. 즉 여자는 작게 굳어진 발을 천으로 싸서절대로 다른 사람에게 보이면 안되었다.

여자를 돈을 주고 사들인 남자는 여자의 어떤 부분보다도 그 숨겨진 발을 보고 애무하는 것이 최고의 감격이었다. 여자의 발을 싸는 신발에 술을 가득 부어 마시는 것이 통달한 사람이라고 했다.

또 어릴때 사들인 여자가 성숙한 처녀가 되면 눈동자를 바늘로 찔러 맹인을 만드는 풍습도 있었다. 이런 여자들은 눈을 뜬 채 맹인이 아닌 눈동자가 깨끗한 여자로 보인다.

다른 남자를 보지 못하게 하는 것과 맹인의 감각이 매우 발달하여 그것이 성에 집약된다는 새리작용을 겨냥한 것이다.

미국

그럼 여기에 등장하는 미국의 여성은 어떤가.

갑자기 벌거벗은 몸을 침대에 던지곤 "헤이, 캄온"(이리오라)...라고 하는 것이 종래부터 전해오는 미국 여성의 침대위의 매너였었다. 이러한 견해는 다분히 일본 남성들의 콤플렉스가 작용한 탓이라 하겠다.

Geld. Guten Tag!" Die Tür fiel ins Schloß. Während Herr Boras die Treppe hinabstieg, überkam ihn abermals das Empfinden, er sei federleicht und schwebe. Auch die Lachlust meldete sich wieder, doch war es eine andere als vorhin.

Auf der Straße—endlich, endlich!—begriff Herr Boras, was geschehen sei: ihm war, kurz gesagt, die Gleichheit mit sich selber abhanden gekommen. Er hatte seine Vergangenheit eingebüßt wie eine Brieftasche, er konnte sich nicht mehr ausweisen. Sonderbar! dachte Herr Boras. Zwar lebe ich, doch scheint es, als hätte ich nie gelebt, denn es sind keine Spuren geblieben. Und dabei war ich von meinem Dasein so fest überzeugt! Nein, es *kann* keine Einbildung gewesen sein. Wie aber habe ich das alles verloren! Vielleicht durch eine ungeschickte Bewegung? Richtig, so wird's sein: ich bin aus dem Weltplan herausgerutscht und passe nun nirgends mehr hinein. Jeder Komet ist planmäßiger als ich.

Inzwischen war es ein Uhr nachmittags geworden. Obwohl Herr Boras, wie er meinte, dem Gefüge der Welt nicht mehr angehörte, spürte er Hunger, denn um diese Zeit pflegte er zu essen—sofern er überhaupt von Gepflogenheiten reden durfte. Er hielt Umschau nach einer Gastwirtschaft, doch damit stand es in dieser Gegend nicht zum besten; der abgelegene Vorort war nur zum Wohnen eingerichtet.

Trübe schritt Herr Boras an vielen Gärten, an vielen Häusern vorbei; manche ähnelten ungemein dem Hause, welches er bislang für das seine gehalten hatte. Deshalb war er auch nicht sonderlich erstaunt, als eine Frau sich aus einem Fenster beugte und ihm zurief: „Zeit, daß du kommst! Die Suppe steht schon auf dem Tisch."

Ohne lange zu überlegen, klinkte Herr Boras die Gartenpforte auf und trat ein; er hatte Hunger. An der Haustür sprang ihm ein Knabe entgegen. „Vati, es gibt Eierkuchen!"

„Fein, mein Junge!" erwiderte Herr Boras. Er streifte den Staub von den Schuhen, hing seinen Hut an den Haken, gab der Frau einen flüchtigen Kuß, setzte sich zu Tisch und begann die Suppe zu löffeln. Während des Essens betrachtete er die Frau und den Jungen, vorsichtig, damit es ihnen nicht auffiel, denn sie hielten ihn offenbar für den Hausvater. Die Frau war nicht übel, und auch der Junge gefiel ihm; das Essen schmeckte gut.

that took place during the 1960's and which were to have a far-reaching impact on the subsequent development and applications of OCR follows.

Throughout the decade, a variety of different detection systems were developed which significantly improved the accuracy of pattern detection. These developments had two major effects upon the OCR industry: first, very reliable readers were manufactured and offered commercially; and, second, readers were built to detect almost any consistently formed character patterns.

Electronic innovations such as the development of transistors allowed less expensive readers to be built. Transistorized readers were more compact and more efficient than their earlier predecessors. Software and hardware refinements were progressively made in the recognition process. The key effect of these refinements was the potential to construct multifont optical readers with the capability of recognizing any of a number of font styles that might appear on documents.

Major improvements and new developments were also made in the essential corollary technology of "transport" design. Readers designed with these new transport features handled pages as large as nine inches by fourteen inches, or documents (i.e. miscellaneous pieces of paper) as small as two inches by three inches. Greater transport capabilities for document-handling attracted many new users with various and different application challenges for OCR systems.

これまでワープロのすぐれたところばかりを述べてきました。しかし、ワープロは何でもやってくれる、といってワープロにすべてをまかせていいというわけではありません。やはり、ワープロはあくまでも機械なのであって、それをどう有効に使うかは、使うひと次第であります。

かな漢字変換方式について考えてみましょう。かなのキーを打ち、変換のキーを押せば、漢字が次次とディスプレーに出てきます。あるいは、かなを打って、辞書（ワープロの漢字辞書）と拡張機能のキーを押せば、漢字がこれまた幾つも並びます。この中からどれを選ぶかは、ワープロを使うひとの日本語の知識によるのです。

例えば、「しそう」と打ったとします。これをワープロは、「死相」という字を出します。これではあんまりだ、というわけで、もう一度キーを押すと「志操」という字が出ます。これも違う。そこでもう一度キーを押す。志操の代わりに「歯槽」と出る。さらにもう一度押す。ようやくにして「思想」が出ました。もし「思想」という言葉を頻繁に用いるのであれば、その次からは「しそう」と打てば、かならず「思想」が第一順位に出てきます。これは、ワープロが記憶した、あるいは学習した、といえるでしょう。

# Scope

- Problem domain
  documents in different languages
  document containing more than one language
- Roman (Latin) based
  European languages
  Swahili, Vietnamese
- Han-based
  Chinese
  Japanese
  Korean
- Partial solution
  Script identification
  Language identification

# Language and Document Images

- Script or character set
  - character forms
- Word shapes
- Writing conventions of the language

# Latin-based Language Classification

- Document sorting by language
- Tuning of character recognition
- Supporting semantic analysis

| Word | German Meaning |
|---|---|
| was | what |
| man | anyone |
| gut | good |
| fast | almost |
| bald | soon |
| tot | dead |

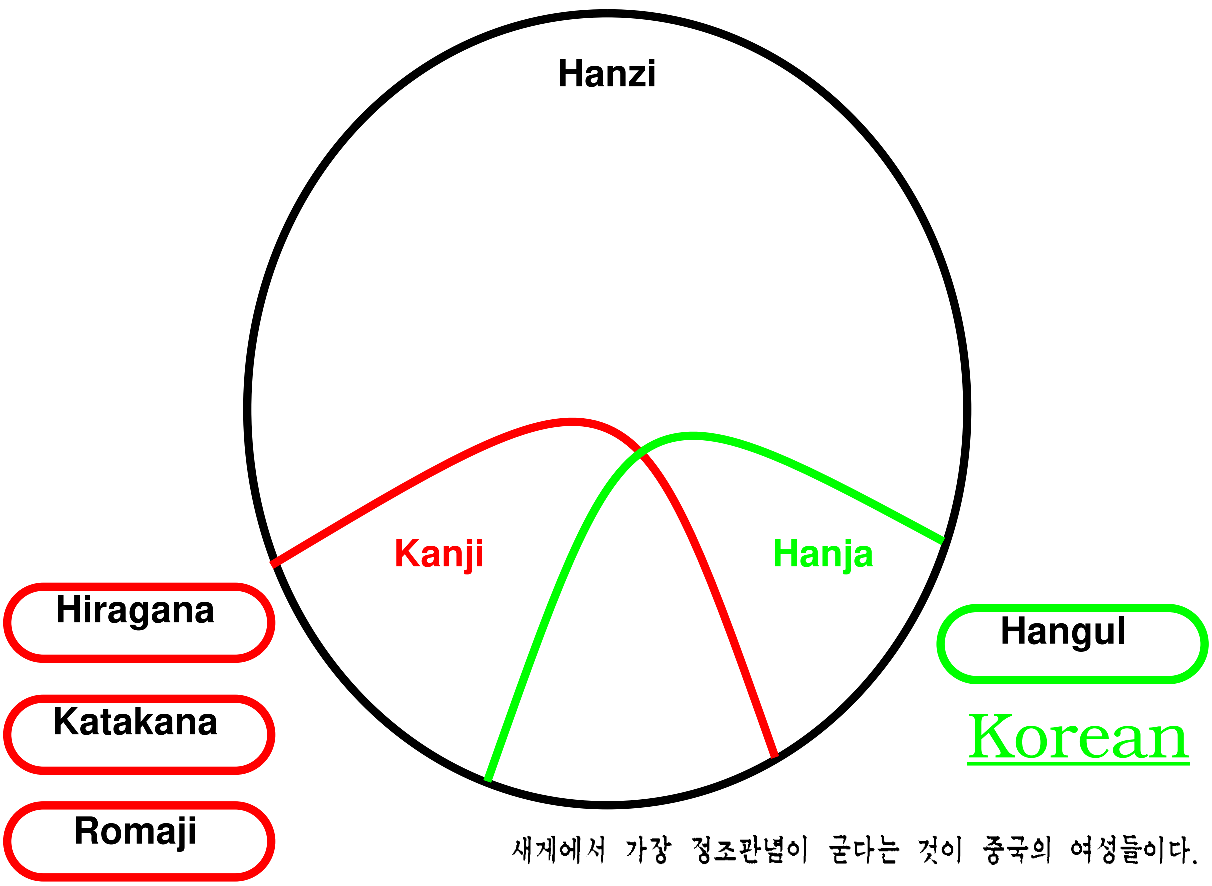
| Word | French Meaning |
|---|---|
| sale | dirty |
| four | oven |
| gave | force feed |
| lest | ballast |
| lice | stadium |
| love | coil (rope) |

# Script

- Roman (or Latin) script is used for many languages
- Serbo-Croatian
- Han script

中國古時候，有一個人姓馬，名字叫書田。

## Chinese




새계에서 가장 정조관념이 굳다는 것이 중국의 여성들이다.

## Japanese

原子核の質量は，原子の質量の大部分を占めるので，

# Character Positioning

- Proportional spacing
- Kerning
  especially prevalent in italic typefaces.
- Mono-spacing
  Chinese and Japanese
  fixed-width characters.
- Justification
  Latin: either ragged-right or right justified
  Han: fully justified

# Text Orientation

- Latin-based horizontal text
- Reading sequence
  - intra-line: left-to-right
  - inter-line: top-to-bottom
  - Exceptions: Arabic and Hebrew
- Asian: either horizontal or vertical
  - horizontal: like Latin-based
  - vertical
    - intra-line: top-to-bottom
    - inter-line: right-to-left
    - Exception: Mongolian

# Word Segmentation

- Latin-based documents
  - words delimited by white space
  - words split across text lines are hyphenated
- Chinese and Japanese documents
  - punctuation indicates word boundary
  - most word boundaries are unmarked
  - semantic and syntactic process
  - no hyphenation.
- Japanese
  - transition: Kana to Kanji
- Korean
  - word spaces
  - no hyphenation

# Character Forms

- Latin-based character
  - diacritics
  - connected component
- Vietnamese
  - lacks f, j, w and z
  - adds seven characters to the (English) Roman character set
  - adds tonal information
  - 18 forms of **a**
- Han-based scripts
  - large numbers of character forms
  - many connected components
- Japanese
  - Kana
  - Furigana
  - Simplified / traditional Chinese
- Korean
  - Hangul
    - 10 basic vowel elements
    - 14 consonantal elements
    - 11,000 possible characters
    - 2300 common use
  - Hanja
    - parenthesized Hangul
- Arabic
  - multiple letter forms

# Capitalization

- Latin-based documents
  - initial words of sentences
  - proper nouns and their associated adjectives
  - German nouns

- Han-based scripts.
  - no analogous rendering

# Symbols

- Quotation marks

| | |
|---|---|
| English | technology of ‘‘transport’’ design. |
| French | à des « connecteurs logiques ». |
| German | „Es war“, sagte mein Onkel, |
| Japanese | その「龍宮城」だが |

- Period, comma

は節形式で記述され，ルールベースに格納されている．判例は「状況と記述」と「双方の主張」の組として，事例ベースに格納されている．概念辞書中の概念の間には，上位/下位の関係に基づく階層構造が成立している（図 2）．概念間の類似性は，この階層構造上の距離で定義される．

新しい事件を事実の集合として表現し，システムに入力すると，二つのエンジンは独立に推論を開始する．ルールベースエンジンの役割はルールベース中のルールを発火して，法的結論を求めることである．しかし，新しい事件の事実集合は通常は具体的な概念(例えば，「運転する」や「殴る」など）のみで記述され，抽象的な概念（例えば「公共の利益」や「基本的人権」など）は含まれない．逆に，法令文には，抽象的な概念が多く含まれている．したがって，初めはルールにマッチするデータが不足し，ルールはほとんど発火せず，データ待ちの状態となる．

一方，事例ベースエンジンの役割は入力された問題と類似する事例（判例）を事例ベースから検索し，検索された事例で使われた原告/被告の主張を流用して，「公共の福祉」などの抽象的な法律概念が成立するための理論を生成することである．原告側の主張を採用すれば，「公共の福祉に反する」ことになり，被告側の主張を利用すれば，「公共の福祉に反しない」を生成する，というように互いに矛盾する結論が生成されることがある．裁判官はどちらの論理を採用するかはわからないので，いずれも仮説である．

事例ベース推論で生成された（仮説としての）法律概念とその理由づけは作業記憶を介してルールベースエンジンに送られ，ルールベースエンジンは推論を再開して，最終的な法的結論が求まる．その際，いろいろな仮説の組合せごとに法令文とつきあわせて，その組合せのもとではどのような法的な結論になるかが求められる．

る地域については記述が乏しいが果してそうであったろうか、とかの疑問が湧いてくるのです。いわば「思いつき」が生まれるのです。

学問において「思いつき」とは重要な役割を演じます。まさにそこに研究者の独創性が現われてくるのです。マックス・ウェーバーの有名な『職業としての学問』は、この「思いつき」の重要さについて語ってくれます。(1)

ウェーバーによると、素人の思いつきには専門家にとって大事なものがしばしばある。しかし、素人が素人であるのは、その思いつきをそのまま放置してしまい、それを具体的な研究に移す手順を知らないからである。これに対して専門家は、思いつきを決まった方法によって仕事の軌道に移すことができる。しかし、思いつきとは机に向かって悩んでいても自ら生まれるものではない。ソファーに寝そべって葉巻をくゆらせている時とか、坂道を散歩している時とかに思いがけなく頭にひらめいてくるものだ。

或る問題について毎日書物を読んでいてもヒントが浮かんでくるものでないことはウェーバーのいう通りです。しかし、坂道を歩きながらいい思いつきが生まれるのは、やはり平常その問題について思いめぐらしているからです。ともかく、或るテーマについての自分の勉学の努力が、いろいろな疑問や新しい着想などを呼び起こすことになるのです。

# Relative character placement

- Character cell
  rectangular / parallelogramic region containing the character

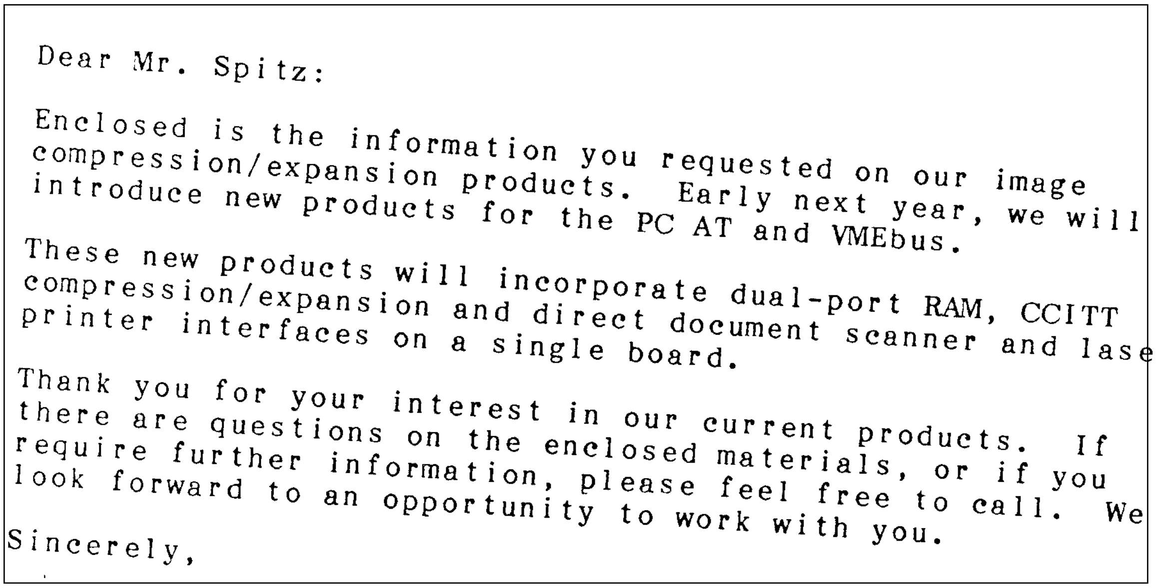

- Arabic, Mongolian, Devanagari have high connectivity
- Ligatures of two or more characters: **ﬁ**, **ﬂ**, **ﬄ**.
- Horizontally disjoint: " 那 에

# Interpolations

- *foreign* language – isolated words or phrases
- Latin-based documents
  typeface change: italic
- Japanese
  script change: Roman, Katakana

# Image processing

## Skew detection



Dear Mr. Spitz:

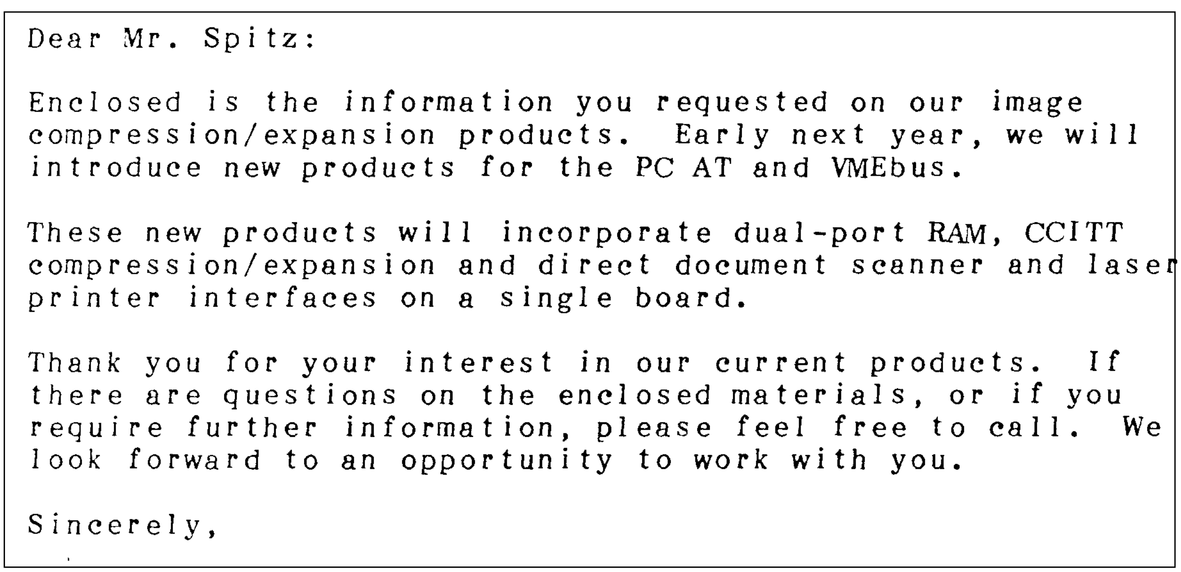
Enclosed is the information you requested on our image compression/expansion products. Early next year, we will introduce new products for the PC AT and VMEbus.

These new products will incorporate dual-port RAM, CCITT compression/expansion and direct document scanner and lase printer interfaces on a single board.

Thank you for your interest in our current products. If there are questions on the enclosed materials, or if you require further information, please feel free to call. We look forward to an opportunity to work with you.

Sincerely,

## Skew compensation



Dear Mr. Spitz:

Enclosed is the information you requested on our image compression/expansion products. Early next year, we will introduce new products for the PC AT and VMEbus.

These new products will incorporate dual-port RAM, CCITT compression/expansion and direct document scanner and laser printer interfaces on a single board.

Thank you for your interest in our current products. If there are questions on the enclosed materials, or if you require further information, please feel free to call. We look forward to an opportunity to work with you.

Sincerely,

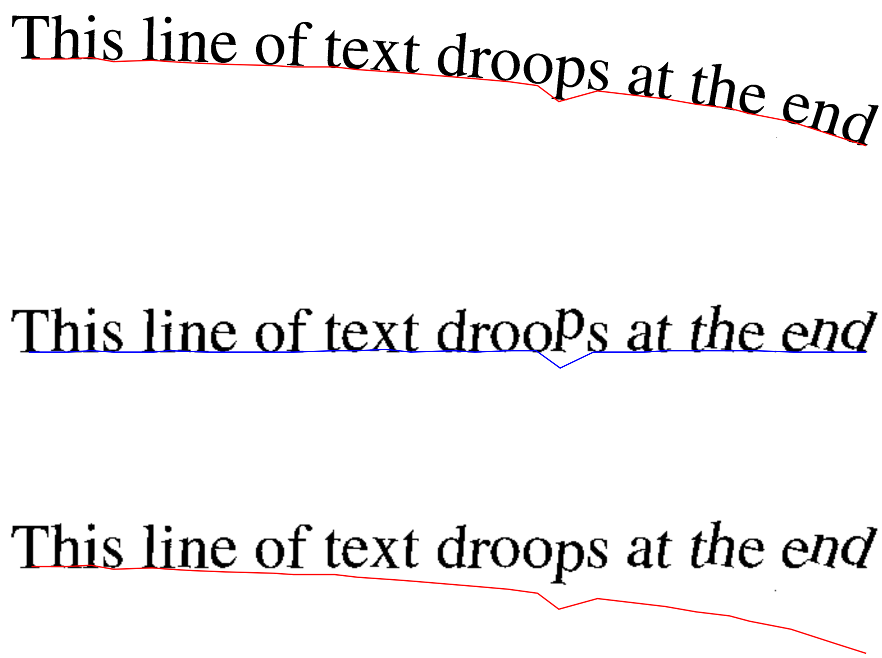
# Text line orientation

原子核の質量は，原子の質量の大部分を占めるので，だいたいその
することになる。したがって，水素原子核の質量を単位として測れば
はほとんどその原子量に等しい値になる。

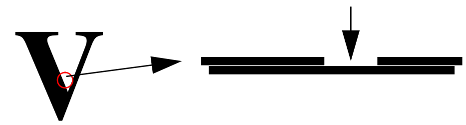
ん、主張したい方を優先したり、整理の仕方を考えることも
「項目内の数値に範囲があるときは、以上・以下・未満な
上・以下・以内・以外・以来・以降・以前・以後」はその数
たとえば、年齢で二十歳以上三〇歳以下を表記したいときは
〔31〜40〕と置きますが、三〇・九歳は当然前者に該当しま
けではありませんから、数値の概念がどの範囲にあるのかと
大きい方から並べることもあります。〔31未満〜20以上〕
〔21〜10〕がくることになります。これは十進法の概念から
ただ、バラツキのくくり方や集計上の考え方にもよりますか

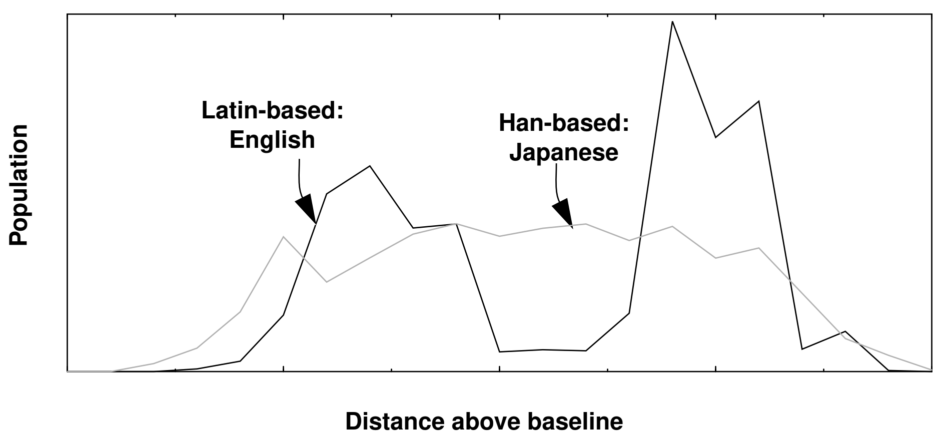
# Text line processing

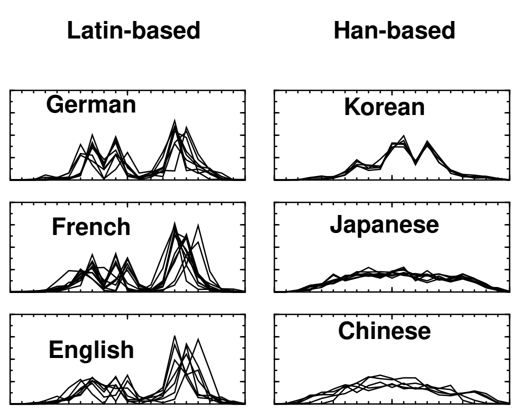
## Line registration



# Script Classification



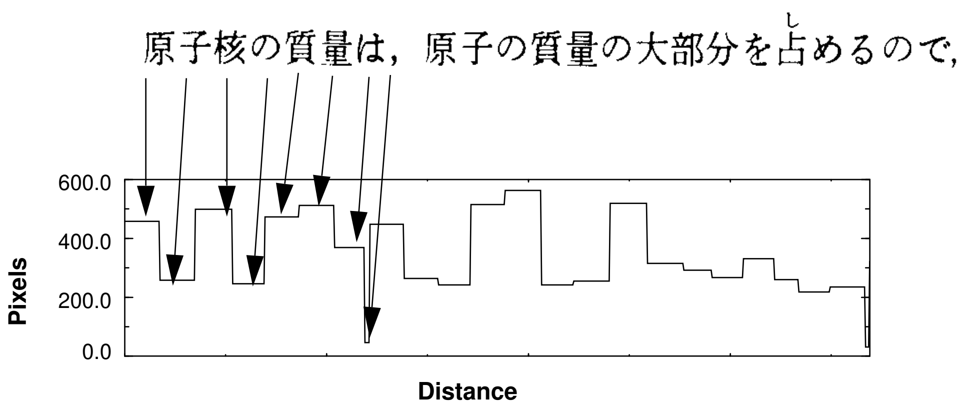



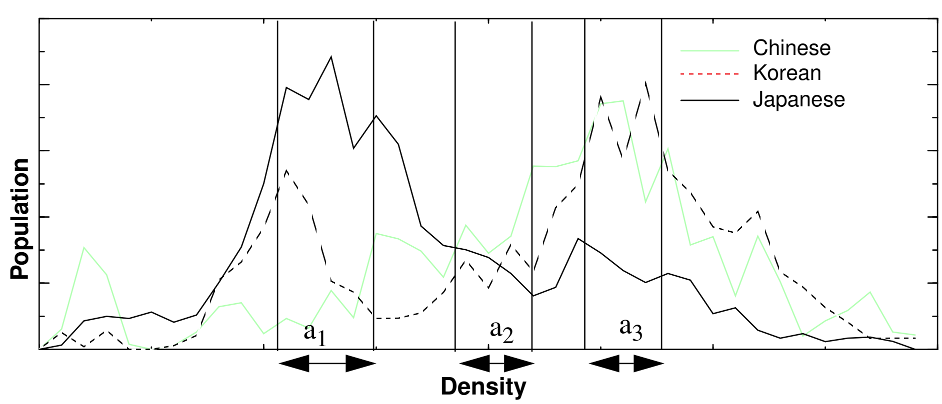
# Han language Classification

原子核の質量は，原子の質量の大部分を占めるので，

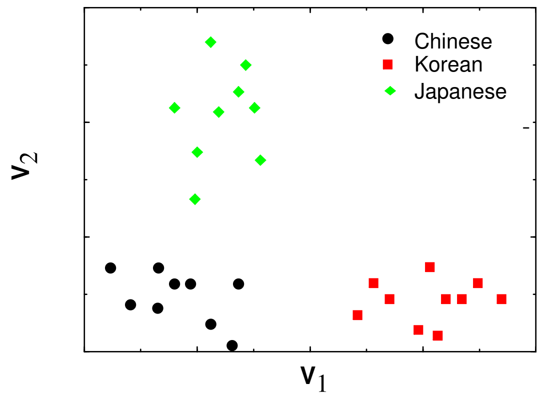
中國古時候，有一個人姓馬，名字叫書田。

새계에서 가장 정조관념이 굳다는 것이 중국의



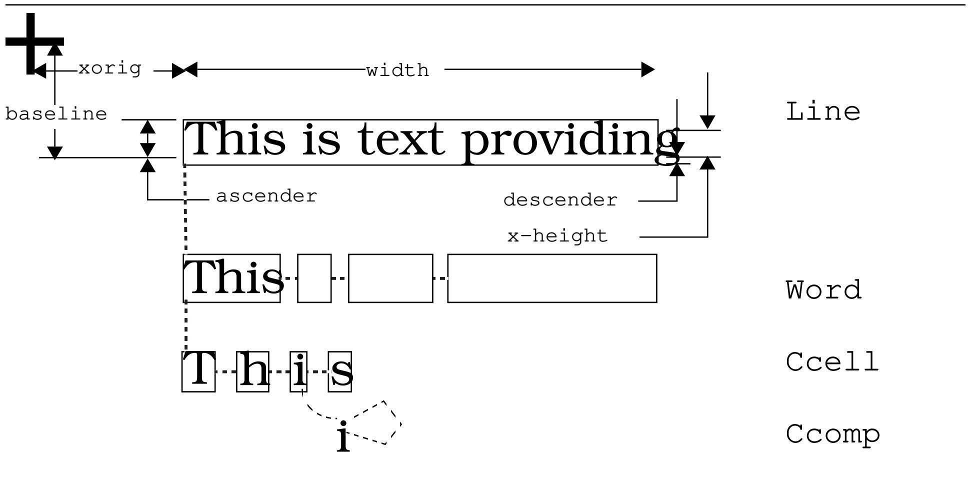
# Classification

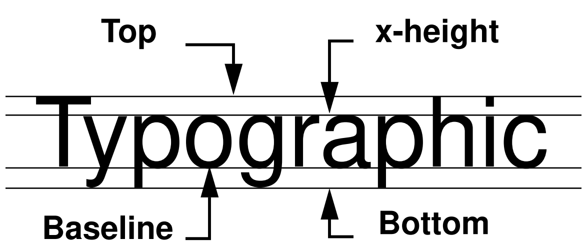



# Accuracy

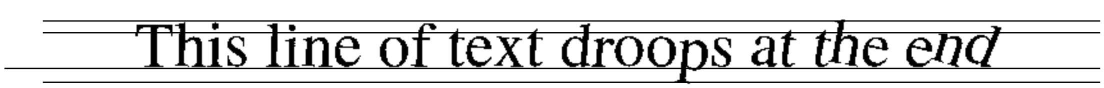
| Lines | Samples | Error rate (%) |
|---|---|---|
| 10 | 72 | 0.0 |
| 8 | 95 | 0.0 |
| 6 | 140 | 0.0 |
| 4 | 217 | 0.5 |
| 2 | 447 | 0.2 |
| 1 | 910 | 15.8 |

# Line Parameters

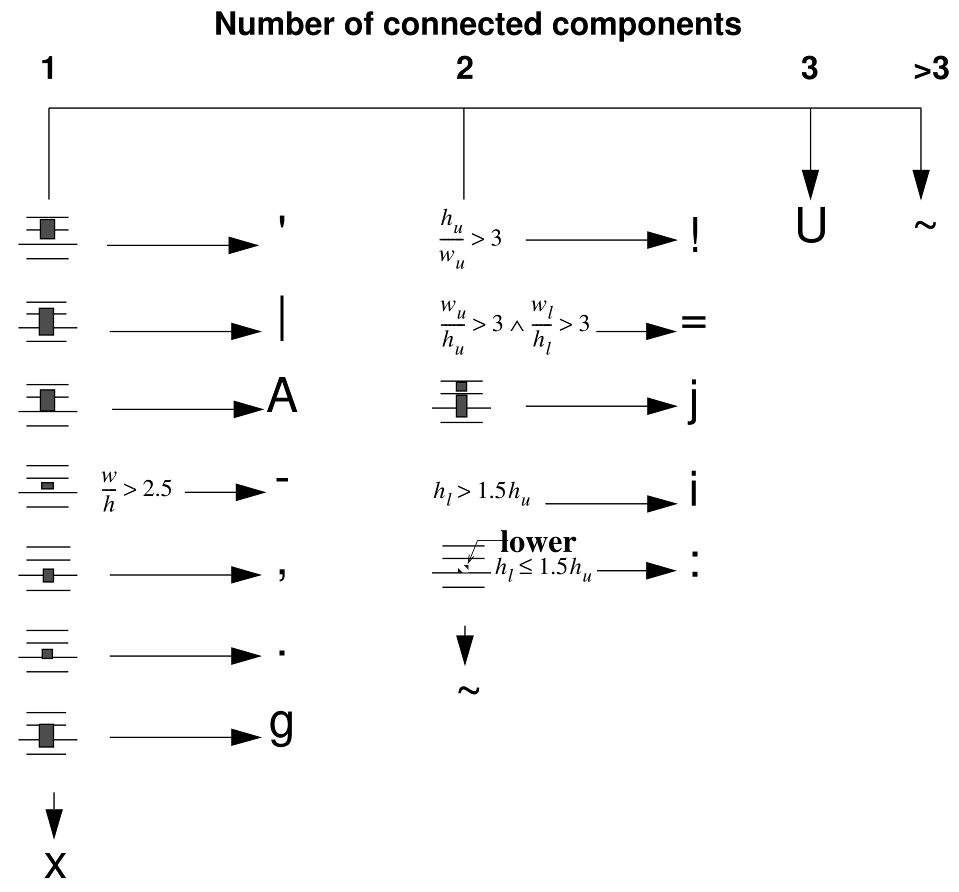



- Parameterization of corrected images



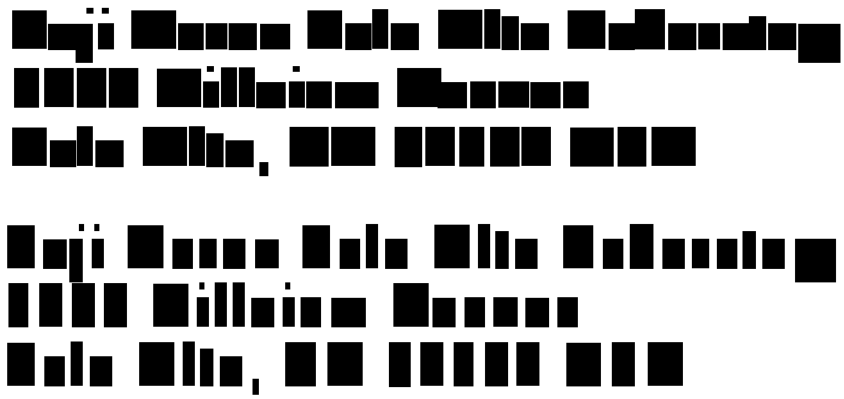
# Character shape / word shape



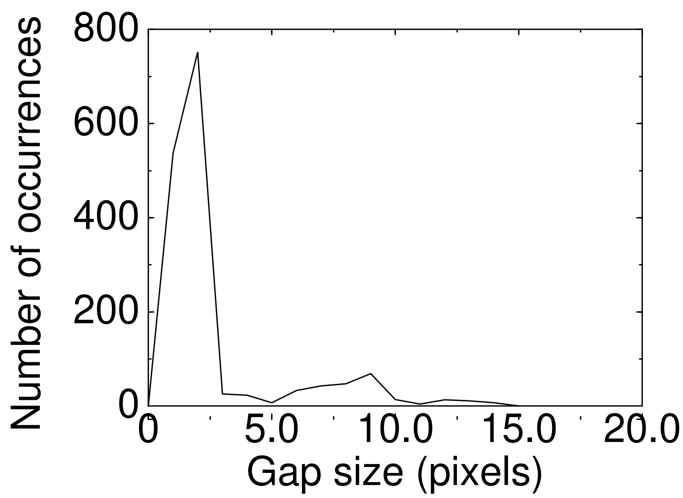

| Character shape code | Character |
|---|---|
| **A** | A-Zbdfhkltß0-9#$&()/[]@{}\| |
| **x** | acemnorsuvwxz |
| **i** | iáàâéèêîôûùñ |
| **g** | gpqyç |
| **j** | j |
| **U** | äëïöüÄÖÜ |

# Word image processing

- Line ends define word boundaries



- Wide gaps define word boundaries




**Character codes**

```
Confidence in the international
monetary system was shaky enough be-
fore last week's action.
```

**Character shape codes**

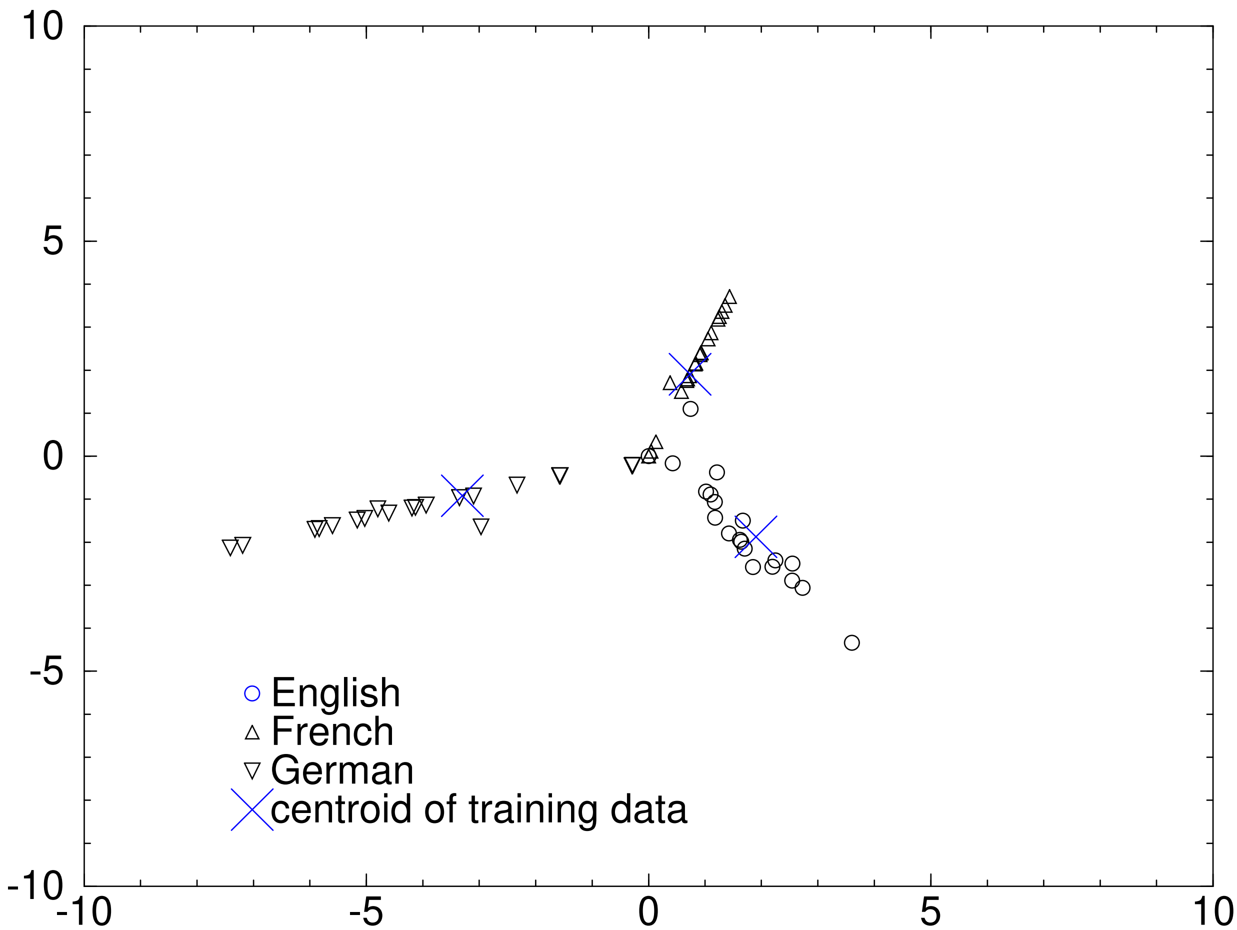
```
AxxAAxxxx ix AAx ixAxxxxAixxxA
xxxxAxxg xgxAxx xxx xAxAg xxxxgA Ax-
Axxx AxxA xxxA'x xxAixxn.
```

# Typesetting effects

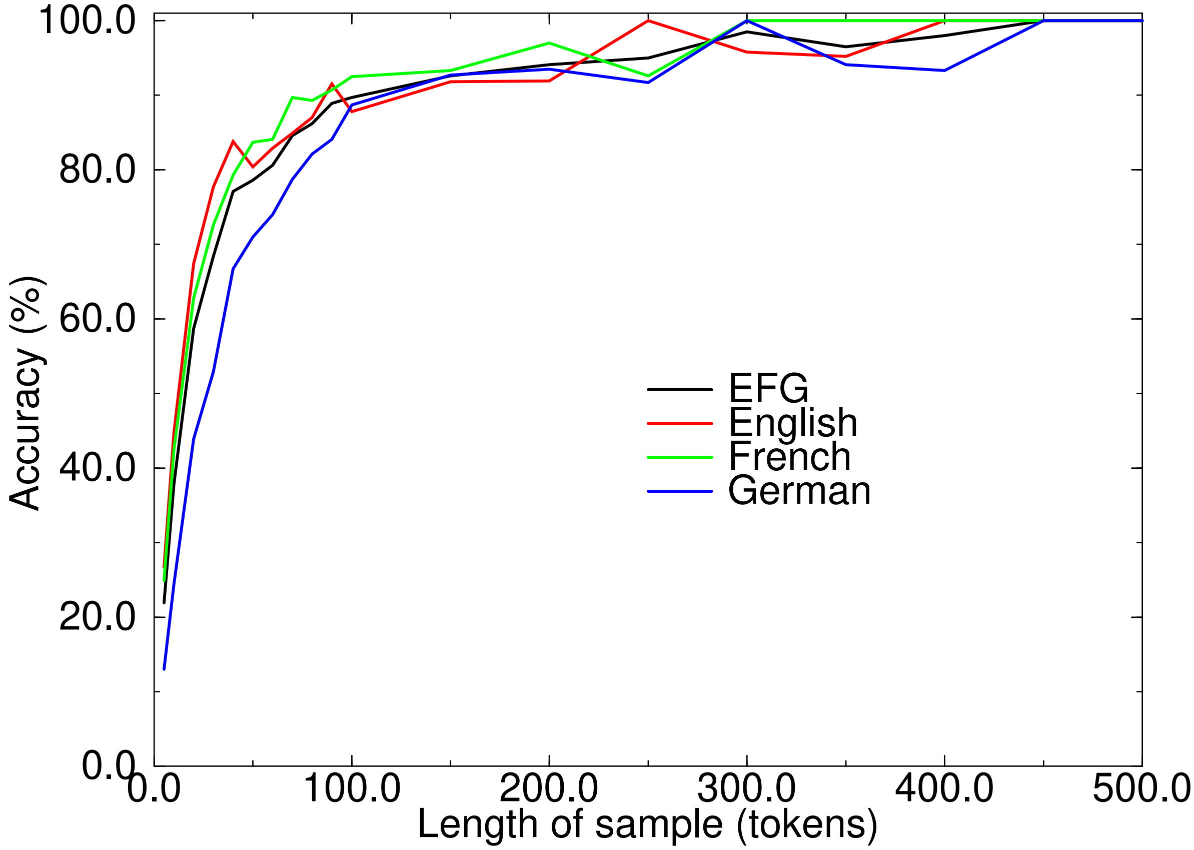
- Ligatures
- fi, ff, fl may be mapped to single CSC
- German vs. ASCII
  ü ue
  ß ss.
  ü maps to **U** and ue to **xx**
- Font dependencies
  è might map to **i** or **A**
- Diacritics on capitals

# Classification

| | English | | French | | German | |
|---|---|---|---|---|---|---|
| **WST** | **Rank** | `Word` | **Rank** | `Word` | **Rank** | `Word` |
| AAx | 1 | `the` | | | | |
| xA | 2 | `of` | 7 | | | |
| Ax | 3 | `to` | 1 | `la le` | 6 | |
| ix | 4 | `is` | | | 10 | |
| xxA | 5 | `and` | | | 3 | `auf` |
| xx | 6 | | 2 | `en` | 2 | `an` |
| Axx | 9 | | 3 | `les` | 1 | `der das` |
| xxx | 8 | | 4 | `aux` | 5 | `wer` |
| gxx | | | 5 | `pas` | | |
| Aix | | | | | 4 | `die` |



# I can name that language in ...



# Automated determination for many languages

- 755 one-page documents in 23 languages

  European languages

  Turkish, Swahili, and Vietnamese.

| Language | Acc (%) | Language type | abbr | Language | Acc (%) | Language type | abbr |
|---|---|---|---|---|---|---|---|
| Afrikaans | 97 | IE | af | Italian | 95 | IE | it |
| Croatian | 100 | IE | cr | Norwegian | 95 | IE | no |
| Czech/Slovak | 44 | IE | cs | Polish | 100 | IE | po |
| Danish | 96 | IE | da | Portuguese | 96 | IE | pt |
| Dutch | 100 | IE | du | Rumanian | 93 | IE | ru |
| English | 95 | IE | en | Spanish | 95 | IE | sp |
| Finnish | 75 | Uralic | fi | Swahili | 97 | | sa |
| French | 92 | IE | fr | Swedish | 98 | IE | sw |
| Gaelic | 86 | IE | ga | Turkish | 93 | | tu |
| German | 97 | IE | ge | Vietnamese | 100 | | vi |
| Hungarian | 94 | Uralic | hu | Welsh | 97 | IE | we |
| Icelandic | 96 | IE | ic | | | | |

- 5 most frequent WSTs from each language
- Total 23 WSTs
- **xx** appears in the top five of 22 languages
- 8 WSTs in the top 5 of exactly one language

  **x**, **xx** , **xxx** , **xxxx** , **i**, **ix** , **xi** , **xix** , **A**, **AAx** , **Ax** , **AxA** , **AxAx** , **Axx** , **Axxx** , **xA** , **xxA** , **Ai** , **Aix** , **g**, **gx** , **xg** , **xxg** , **jx**

# Accuracy

| | Detected Language | | | | | | | | | | | | | | | | | | | | | | | Error Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | en | ge | du | af | fr | it | sp | pt | ru | da | no | se | ic | ga | we | cr | cs | po | hu | fi | tu | sa | vi | |
| en | 36 | 1 | | | 1 | | | | | | | | | | | | | | | | | | | 2 |
| ge | | 29 | | | | | | | | | | | | 1 | | | | | | | | | | 1 |
| du | | | 28 | | | | | | | | | | | | | | | | | | | | | 0 |
| af | | | 1 | 29 | | | | | | | | | | | | | | | | | | | | 1 |
| fr | | | | | 23 | | | 1 | | | | | | 1 | | | | | | | | | | 2 |
| it | | | | | | 35 | | 1 | 1 | | | | | | | | | | | | | | | 2 |
| sp | | | | | | | 39 | 2 | | | | | | | | | | | | | | | | 2 |
| pt | | | | | | | 1 | 25 | | | | | | | | | | | | | | | | 1 |
| ru | | | | | | 3 | | | 38 | | | | | | | | | | | | | | | 3 |
| da | | | | | 1 | | | | | 25 | | | | | | | | | | | | | | 1 |
| no | | | | | | | | | | | 39 | | 2 | | | | | | | | | | | 2 |
| se | | | | | | | | | | | | 40 | | | | | | | | 1 | | | | 1 |
| ic | | | | | | | | | | | | | 23 | | | | | | | 1 | | | | 1 |
| ga | | | | | | 2 | | | 2 | | | | 1 | 31 | | | | | | | | | | 5 |
| we | | | | | | | | | | | | | | | 30 | | | | | | | 1 | | 1 |
| cr | | | | | | | | | | | | | | | | 33 | | | | | | | | 0 |
| cs | | | | | | | | | 1 | | | | | | | 6 | 24 | 16 | 3 | 5 | | | | 31 |
| po | | | | | | | | | | | | | | | | | | 28 | | | | | | 0 |
| hu | | | | | | | | | | | | | | | | | | 2 | 29 | | | | | 2 |
| fi | | | | | | | | | | | | | | | | | 6 | | 2 | 24 | | | | 8 |
| tu | | | | | | | 1 | | | | | | | | | | | | | 1 | 28 | | | 2 |
| sa | | | | | | | | | | | | | | | | | | | | | 1 | 37 | | 1 |
| vi | | | | | | | | | | | | | | | | | | | | | | | 13 | 0 |
| | 0 | 1 | 1 | 0 | 2 | 5 | 2 | 4 | 4 | 0 | 0 | 0 | 3 | 2 | 0 | 6 | 6 | 18 | 3 | 9 | 1 | 1 | 0 | 69 |

# Optical Character Recognition

- Research OCR tuned to specific applications
  multilingual capability has not received much attention
  Baird English, Greek, Tibetan and Japanese (and Chess)
- OCR requires prior knowledge of language
  character set
  post-recognition spelling checker
- Most designed for English text
  other languages in Roman
  document granularity
- Japanese OCR
  Romaji: proportional spacing causes problems
- Complexity: enemy of robustness
- Small morphological differences in character forms
- Computational efficiency
  $F = C - 1$ binary features to resolve $C$ classes.
  memory requirements $O(CF)$
  computational time $O(CF + log\ C)$
- English 90 or more symbols
- Chinese 8000 or more symbols

# Language Models Applied to OCR

- Improve speed / accuracy
- Character n-gram frequencies
  - detect the most likely characters first
  - Bigrams or trigrams recognition progresses
- Post-processing phase “spelling dictionary”
- Word frequency
  - documents contain few distinct words frequently repeated
- Character n-grams for Chinese / Japanese OCR

# Character Codes

- ASCII is adequate for English

  coöperate

- ISO Latin-1 (ISO-8859-1)

  does not cover Polish, Czech, Hungarian, etc.

- ISO-8859-n

- Unicode

| | **Language** | **Standard** | **Code** |
|---|---|---|---|
| 生 | Japanese | JIS X0208-1983 | 3224 |
| | Simplified Chinese | GB 2312-80 | 4190 |
| | Traditional Chinese | CNS 11643-86 | A5CD |
| | Korean | KS C5601-1987 | 6370 |
| | All | Unicode | 751F |

# Applications

- Cross-lingual information retrieval
  - patents
  - business news
  - regulatory information
- Document sorting
  - character recognition
  - translation
  - word spotting

# Possible Extensions

- Breadth
  Cyrillic, Greek or Hebrew scripts
  Arabic, Devanagari
- Depth
  Cross-lingual access
  Document understanding

# Acknowledgments

Joseph Becker
Jean-Marie de La Beaujardière
Arlene Holloway
David Hull
Takehiro Nakayama
Masaharu Ozaki
Penelope Sibun
Karl Tombre