かな漢字変換の仕組み

@xmmm

本題に入る前に…

自己紹介

twitter id: @xmmm

苫小牧高専電気電子工学科
→ 千葉大学工学部
→ 東京大学大学院工学系研究科

どう見てもロンダです．本当に (ry

かな漢字変換って？

日本語文の読み

↓

かな漢字混じり文

にわにはにわにわとりがいる

庭には二羽鶏がいる

MS-IME, ATOK, ことえり , Canna, Wnn, Prime, Anthy, SKK, Social IME, etc...

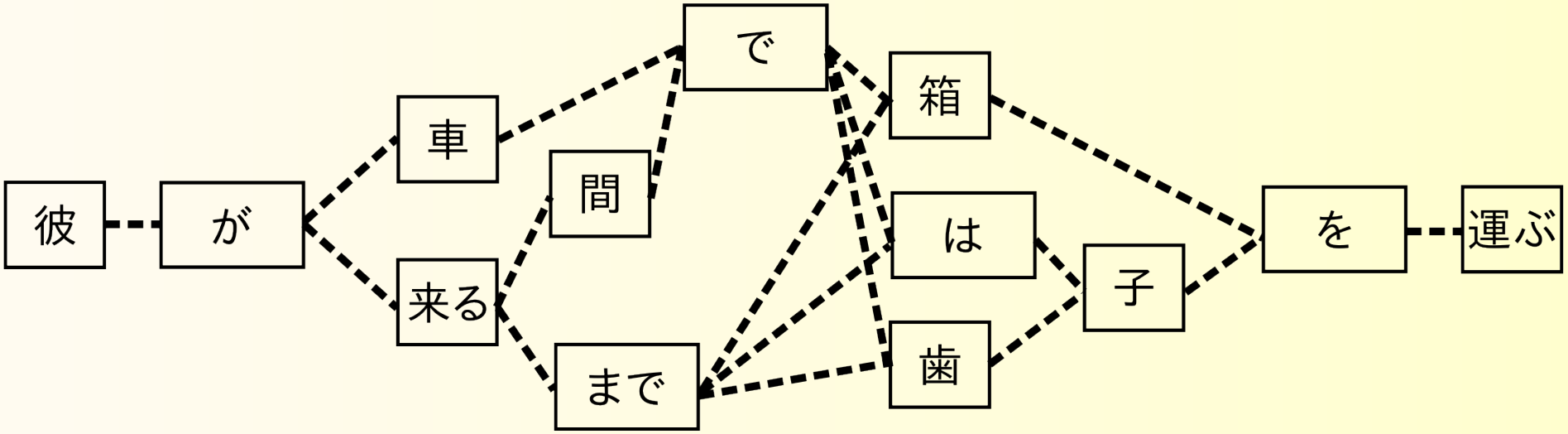
かな漢字変換の手順

二段階で変換する

1. 辞書検索
2. 最適解の選択

1. 辞書検索

「かれがくるまではこをはこぶ」



辞書を引いて「形態素」のラティスを作る

2. 最適解の選択

「彼が車で箱を運ぶ」

「彼が来るまで箱を運ぶ」

「彼が来るまでは子を運ぶ」

etc...

どの文（どのルート）を選ぶ？

選択の仕方

a)左最長一致法

b)2 文節最長一致法（ATOK)

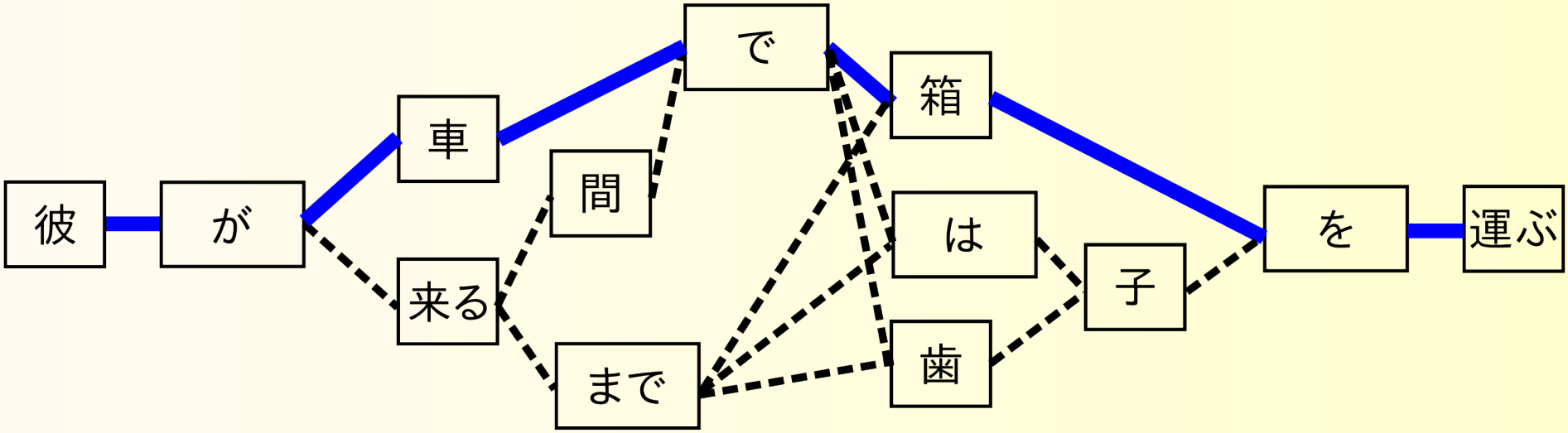
c)文節数最小法（Wnn の一部)

d)コスト最小法（MS-IME，ことえり)

e)確率モデル（Anthy)

a) 左最長一致法

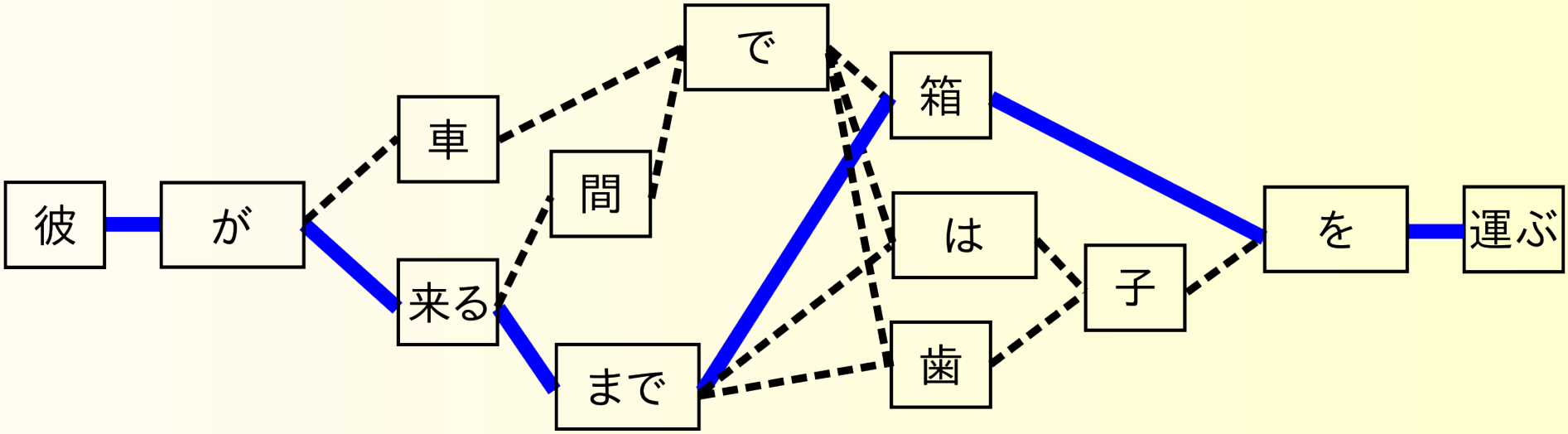
左側から「長い」形態素を選択する



カンタンだけど精度が良くない

b) 2 文節最長一致法 (ATOK)

連続 2 文節が最長となるものを選択



処理コストはそこそこ

b) 2 文節最長一致法 (ATOK)

連続 2 文節の長さが同じときは…

#1　くる / まで

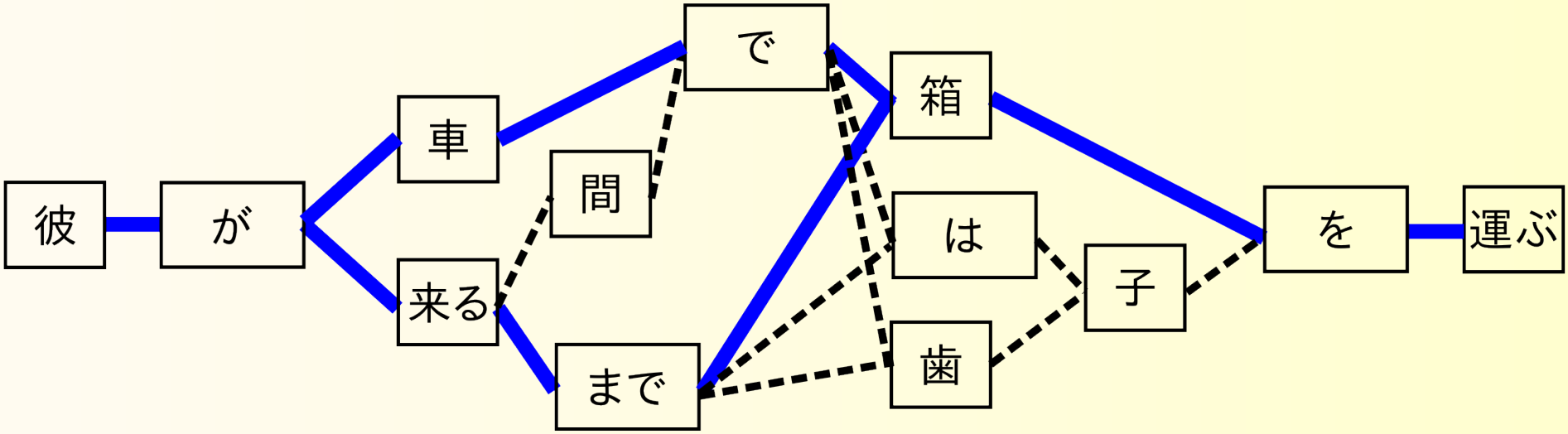
#2　くるま / で

前方の文節が短い #1 を採用

(なぜか分からないけどそうすると上手くいく)

c）文節数最小法（Wnn の一部）

文節の数が最少になるように選択



処理コストは高いが正解率も高い

一意に決定できないこともしばしば

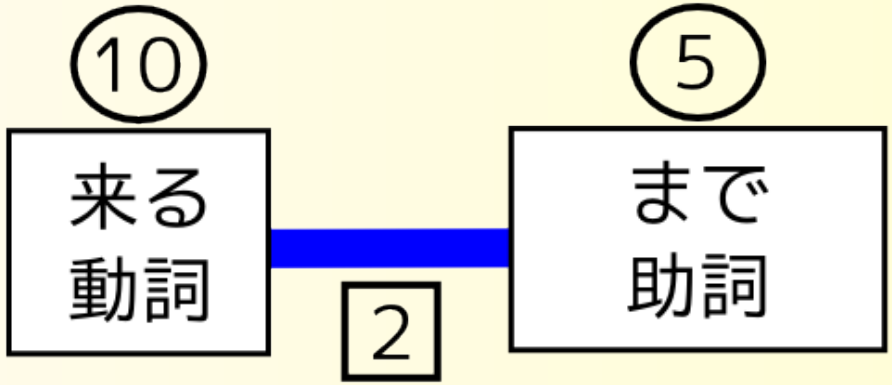
d) コスト最小法 (MS-IME, ことえり)

・形態素コスト…形態素・品詞間のコスト

e.g. ' くる '-[動詞]…10pt

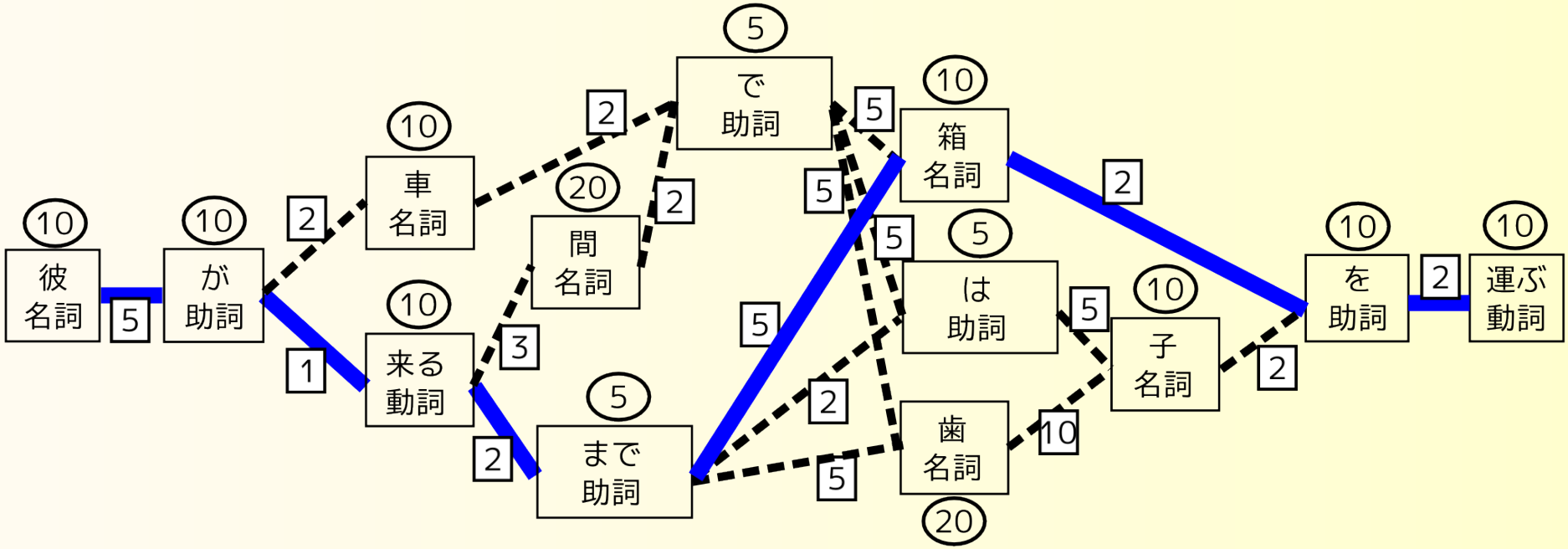
・連接コスト…品詞間のコスト

e.g. [動詞]-[助詞]…2pt



d) コスト最小法 (MS-IME, ことえり)

コストが最小となるルートを選択



計算がとてもメンドウ
コストを人手で生成

e) 確率モデル(Anthy)

コスト最小法を確率により表現

数学的な裏付けのある変換結果

Anthy…2002 年未踏ソフトウェア

e) 確率モデル(Anthy)

(ry

時間が足りません＞＜

ググる：「HMM　形態素解析」

「統計的かな漢字変換」

『隠れマルコフモデルによる日本語形態素解析のパラメータ推定』

http://ci.nii.ac.jp/naid/110002721502/

同音語の判別

これまでの方法だけでは上手くいかない

同音異義語

自己 / 事故

厚い / 暑い / 熱い

→ 「共起語辞書」を使う

共起語辞書

２単語をペアにしたデータ

例えば…

（本，厚い），（お湯，熱い），

（夏，暑い），（鉄，熱い）

「お湯が｛暑い，熱い，厚い｝」

➡ 「お湯が熱い」

じゃあどの IME がいいの…？

ぶっちゃけ好みの問題

勝手に skk をお勧めしてみる

skk

shift キーを使って区切りを明示

”Gohann[SPACE]TaBetai”

「御飯食べたい」

skk

区切りを明示できるかわりに

かなり変態な打ち方を強要される

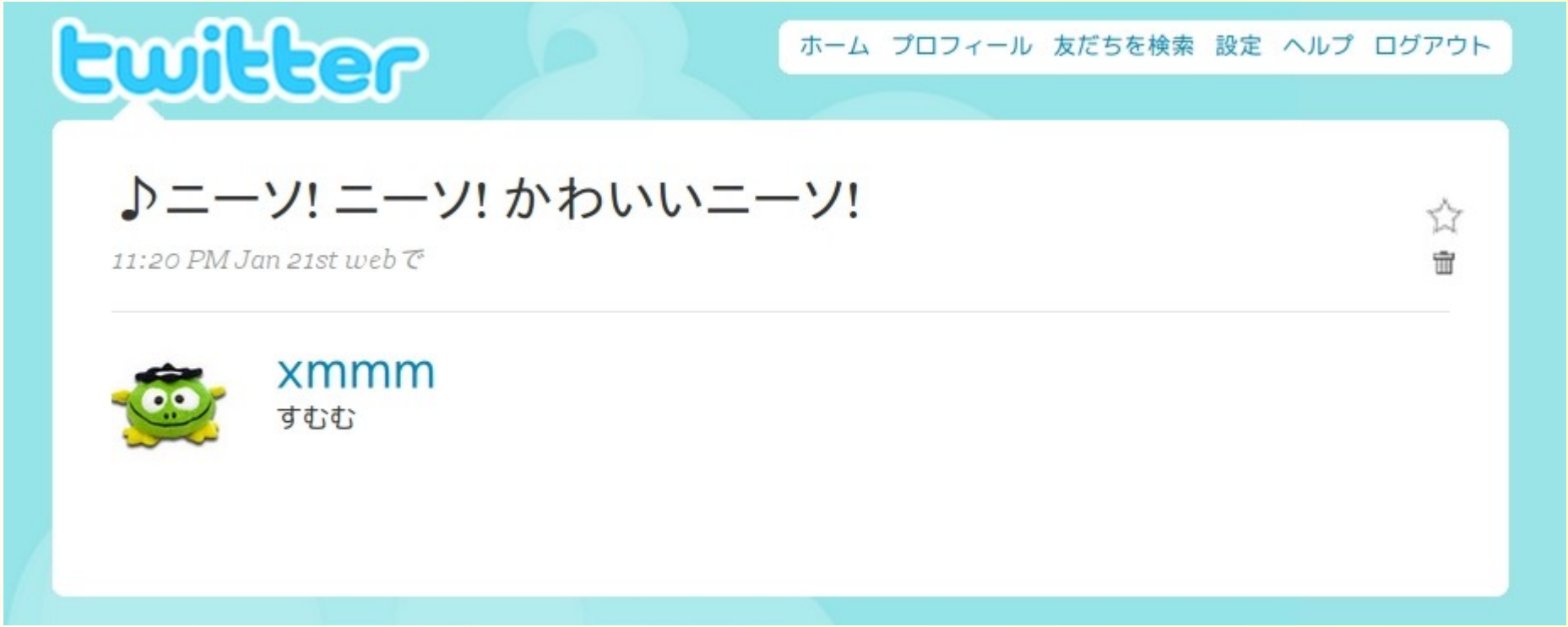
変態に定評のある高専生

↓

skk は高専生向き…？

skk

こんな文でも誤変換が少ない！



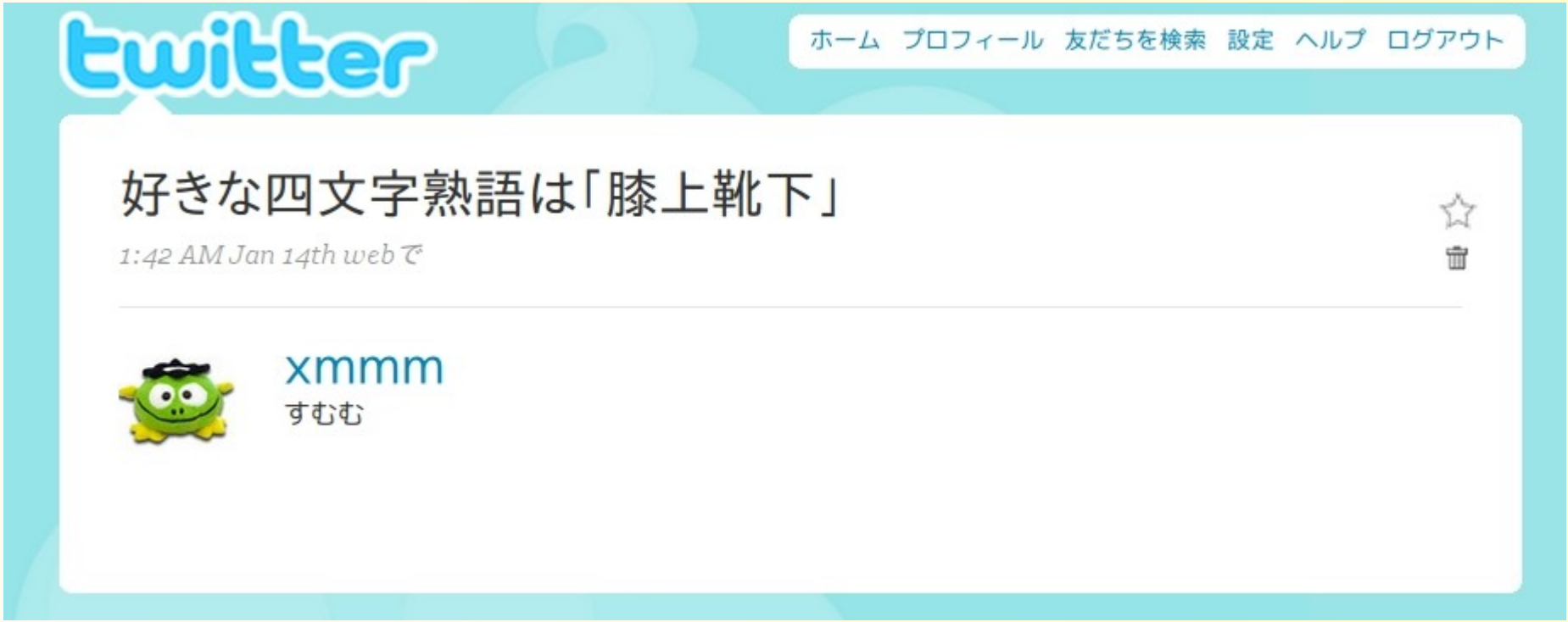
skk

こんな文でも！


twitter
ホーム プロフィール 友だちを検索 設定 ヘルプ ログアウト
そのニーソが いいねと僕が 言ったから 1月11日は
ニーソ記念日
10:15 PM Jan 1st pochitterで
xmmm
すむむ

skk

こんな文でも！！


twitter
ホーム プロフィール 友だちを検索 設定 ヘルプ ログアウト
好きな四文字熟語は「膝上靴下」
1:42 AM Jan 14th web で
xmmm
すむむ

skk

こんな文でも！！！


twitter
ホーム プロフィール 友だちを検索 設定 ヘルプ ログアウト
ニーソが好きで何が悪い!
12:50 AM Jan 22nd web で
xmmm
すむむ

skk

こんな文でも！！！！


twitter
ホーム プロフィール 友だちを検索 設定 ヘルプ ログアウト
お知らせ：彼女が出来ました。Skypeなう。
2:05 AM Aug 1st NatsuLiphone で
xmmm
すむむ

skk

変換できます！

最後に

- 意外とヒューリスティック
- かな漢字変換のこれからをお楽しみに！

時間が余ったようなので…

どれを使っていますか？

- MS-IME
- ATOK
- ことえり
- Cannna
- Wnn
- Anthy
- Prime
- SKK