# Uči (うち、ウチ、内)

Prvním vyšetřeným pomocným podstatným jménem je uči. Jeho hlavním významem je *uvnitř*, ale objevuje se i v konstrukcích s významem posunutým na časové užití, které lze přeložit jako *dokud*. Protože upřesňuje časoprostorový vztah svého přívlastku, jedná se pomocné jméno s předložkovým užitím. S mírou gramatikalizace jedenáct bylo zvoleno jako více gramatikalizovaný zástupce své kategorie.

## První náhled

Po stažení dat a jejich filtraci zůstalo 33 199 záznamů, z čehož 30 650 (92,32%) je zapsáno hiraganou, 87 (0,26%) katakanou a zbylých 2 462 (7,42%) znakem. Podrobnosti jsou vidět v následujících tabulkách:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Subkorpus** | **ウチ** | **うち** | **内** | **Součet** |
| **Vydané texty** | **20,69%** | **32,33%** | **36,31%** | **32,59%** |
| Vydané texty - noviny | 0,00% | 3,66% | 1,68% | 3,49% |
| Vydané texty - knihy | 77,78% | 86,78% | 92,39% | 87,23% |
| Vydané texty - časopisy | 22,22% | 9,56% | 5,93% | 9,28% |
| **Knihy v knihovnách** | **4,60%** | **33,39%** | **34,48%** | **33,40%** |
| **Ostatní texty** | **74,71%** | **34,28%** | **29,20%** | **34,01%** |
| Ostatní - blogy | 76,92% | 17,48% | 35,33% | 18,96% |
| Ostatní - bestselery | 3,08% | 12,84% | 18,22% | 13,12% |
| Ostatní - vládní jednání | 0,00% | 9,22% | 1,11% | 8,65% |
| Ostatní - informační bulletiny | 1,54% | 5,41% | 3,34% | 5,25% |
| Ostatní - učebnice | 0,00% | 2,22% | 1,39% | 2,15% |
| Ostatní - zákoníky | 0,00% | 6,38% | 0,70% | 5,98% |
| Ostatní - vládní zprávy | 0,00% | 27,89% | 3,89% | 26,20% |
| Ostatní - Čiebukuro | 18,46% | 18,01% | 35,33% | 19,11% |
| Ostatní - verše | 0,00% | 0,56% | 0,70% | 0,57% |

Tabulka 1: Distribuce uči v korpusu (NINJAL 2009)

V *Tabulce 1* jsou zaneseny výskyty slova uči v korpusu a zobrazeny jako procenta všech výskytů tohoto slova. Jednotlivé způsoby zápisu jsou rozděleny po sloupcích a ve sloupci Součet jsou počty výskytů uči v daném subkorpusu sečteny. Zvýrazněny jsou tři hlavní subkorpusy. Za zmínku stojí hodnoty sloupce Součet těchto tří záznamů. Výskyty uči jsou v těchto třech subkorpusech rozděleny zhruba po třetinách (32,59%, 33,40% a 34,01%).

Při jemnějším dělení těchto subkorpusů se získá 13 různých subkorpusů nižší úrovně, které jsou zastoupeny nezvýrazněnými řádky. Procentuální údaje v těchto řádcích udávají poměrné zastoupení vůči nadřazenému subkorpusu. Například z tabulky lze zjistit, že v subkorpusu *vydaných textů* se uči psané katakanou objevilo v 20,69% všech výskytů, z čehož 77,78% výskytů bylo v *knihách* a 22,22% v *časopisech*. Subkorpus *knih v knihovnách* žádné další dělení nemá. Zde je hoden zvláštního zřetele fakt, že většina výskytů zápisu katakanou pochází ze subkorpusu *ostatních textů* (74,71%), konkrétně z *blogů* (76,92%) a *Čiebukura* (japonská verze Yahoo! Answers, 18,46%), což jsou internetové zdroje. Tyto dva subkorpusy do své skupiny také přispěly většinovým podílem zapisů znakem (v součtu 70,65%).

Tento druh tabulky je vhodný pro první náhled na data a utvoření si základního přehledu, pro podrobnější interpretaci ale už vhodný není, protože nezohledňuje velikosti jednotlivých subkorpusů. Mohlo by se zdát, že například *vydané knihy* mají oproti *časopisům* a *novinám* vysoký podíl zápisu uči hiraganou (86,78%), ale to je způsobeno tím, že v tomto subkorpusu je výskyt uči vysoký celkově (87,23%). Není překvapivé, že když se druhé dva subkorpusy dělí o zbývajících 13,22% výskytů uči v hiraganě, budou jejich hodnoty oproti *knihám* menší.

*Tabulka 2* ukazuje procentuální zastoupení způsobů zápisu pro každý subkorpus. Zde je už každý řádek nezávislý na ostatních, nezvýrazněné řádky se interpretují stejně jako zvýrazněné. Řádek Celý korpus obsahuje průměrné podíly zápisů pro celý korpus. V této tabulce zaujmou velmi nízké podíly zápisu znakem v *přepisech vládních jednání*, *zákoníků* a *vládních zpráv*. Společnou vlastností těchto subkorpusů je to, že všechny obsahují texty oficiální a formální.

Další náhled na data je v *Tabulce 3*. V této tabulce jsou relativní četnosti různých zápisů slova uči na milion slov. Tyto relativní četnosti se získají vydělením počtu výskytů zkoumaného slova počtem všech slov v subkorpusu a vynásobením vzniklé hodnoty milionem. (Brezina 2018) Toto umožňuje porovnávat jednotlivé subkorpusy přímo mezi sebou, protože je odstraněna závislost na velikosti korpusu. Je pozoruhodné, že se v subkorpusech *Ostatní - zákoníky* a *Ostatní - vládní zprávy* uči vyskytuje mnohem hustěji než v ostatních.

|  |  |  |  |
| --- | --- | --- | --- |
| **Subkorpus** | **ウチ** | **うち** | **内** |
| **Vydané texty** | **0,17%** | **91,57%** | **8,26%** |
| Vydané texty - noviny | 0,00% | 96,03% | 3,97% |
| Vydané texty - knihy | 0,15% | 91,10% | 8,75% |
| Vydané texty - časopisy | 0,40% | 94,32% | 5,28% |
| **Knihy v knihovnách** | **0,04%** | **92,31%** | **7,66%** |
| **Ostatní texty** | **0,58%** | **93,06%** | **6,37%** |
| Ostatní - blogy | 2,34% | 85,80% | 11,86% |
| Ostatní - bestselery | 0,13% | 91,03% | 8,84% |
| Ostatní - vládní jednání | 0,00% | 99,18% | 0,82% |
| Ostatní - informační bulletiny | 0,17% | 95,78% | 4,05% |
| Ostatní - učebnice | 0,00% | 95,88% | 4,12% |
| Ostatní - zákoníky | 0,00% | 99,26% | 0,74% |
| Ostatní - vládní zprávy | 0,00% | 99,05% | 0,95% |
| Ostatní - Čiebukuro | 0,56% | 87,67% | 11,77% |
| Ostatní - verše | 0,00% | 92,19% | 7,81% |
| **Celý korpus** | 0,26% | 92,32% | 7,42% |

Tabulka 2: Podíly způsobů zápisu v jednotlivých subkorpusech (NINJAL 2009)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Subkorpus** | **ウチ** | **うち** | **内** | **Součet** |
| **Vydané texty** | **0,53** | **289,33** | **26,11** | **315,96** |
| Vydané texty - noviny | 0,00 | 265,01 | 10,95 | 275,96 |
| Vydané texty - knihy | 0,49 | 302,21 | 29,03 | 331,73 |
| Vydané texty - časopisy | 0,90 | 214,03 | 11,98 | 226,91 |
| **Knihy v knihovnách** | **0,13** | **337,67** | **28,01** | **365,82** |
| **Ostatní** | **1,62** | **262,31** | **17,95** | **281,88** |
| Ostatní - blogy | 4,94 | 181,42 | 25,08 | 211,44 |
| Ostatní - bestselery | 0,54 | 360,92 | 35,05 | 396,50 |
| Ostatní - vládní jednání | 0,00 | 189,91 | 1,57 | 191,48 |
| Ostatní - informační bulletiny | 0,27 | 151,45 | 6,40 | 158,11 |
| Ostatní - učebnice | 0,00 | 251,91 | 10,81 | 262,72 |
| Ostatní - zákoníky | 0,00 | 620,90 | 4,63 | 625,53 |
| Ostatní - vládní zprávy | 0,00 | 600,50 | 5,74 | 606,24 |
| Ostatní - Čiebukuro | 1,17 | 184,85 | 24,82 | 210,84 |
| Ostatní - verše | 0,00 | 264,36 | 22,40 | 286,76 |
| **Celý korpus** | 0,83 | 292,99 | 23,53 | 317,35 |

Tabulka 3: Počty výskytů slova uči na milion slov (NINJAL 2009)

## Logistická regresní analýza

Předcházející tabulky nabídly několik různých náhledů na výskyty pomocného podstatného jména uči v korpusu. Tyto náhledy od sebe vždy odlišovaly jednotlivé subkorpusy. Nabízí se tedy otázka, jestli pozorované rozdíly jsou statisticky významné a jestli má rozdělení těchto výskytů do subkorpusů na vyskytující se způsoby zápisu nějaký vliv, a pokud ano, jaký.

Před použitím regresní analýzy je třeba udělat několik rozhodnutí. Pro každou z proměnných, které se v ní vyskytnou, je třeba vybrat jednu hodnotu, která bude prohlášena za základní a vůči ní porovnány všechny ostatní hodnoty dané proměnné. Ve vyšetření vlivu subkorpusů na způsob zápisu vystupují proměnné dvě – příslušnost k subkorpusu a způsob zápisu. Způsob zápisu může nabývat dvou hodnot. Zápis je zapsán buď kanou, nebo znakem. Použije-li se za základní hodnotu zápis kanou, ve výsledné tabulce se objeví čísla, která se interpretují jako šance na zápis znakem. Při použití znaku jako základní hodnoty vznikne naopak šance na zápis kanou. Obě možnosti si jsou ekvivalentní. Jako základní hodnota byl pro tuto práci zvolen zápis znakem. Interpretace výsledku „šance na zápis kanou“ je vhodnější pro ta pomocná jména, pro která je zápis kanou dominantním způsobem zápisu a těchto je mezi zkoumanými jmény víc. Je sice možné volit tuto hodnotu jinak pro každé zkoumané pomocné jméno, ale v zájmu snadné porovnatelnosti výsledků je lepší, jsou-li základní hodnoty zvoleny u všech jmen stejně. Druhou proměnnou je příslušnost k subkorpusu, kde byl za základní zvolen subkorpus *Vydané texty - knihy*. Tento subkorpus se ukázal jako velmi dobrý základ pro porovnávání, protože všechna zkoumaná pomocná jména v něm dosahují svých průměrných hodnot a to i v četnosti, i v podílu jednotlivých způsobů zápisu.

Výsledkem této analýzy je model v *Tabulce 4*. Tato tabulka je převzata z online nástroje Lancaster Stat Tools online (Brezina 2018) a skládá se ze dvou částí. První částí je popis modelu. Prvním údajem v tomto popisu je výsledek testu pomocí poměru věrohodností, který má za cíl ověřit, že je tento model statisticky významně lepší, než by byl model bez vysvětlujících proměnných. (Brezina 2018) Pokud by v tomto testu vyšlo, že rozdíl není statisticky významný, znamenalo by to, že v další interpretaci by už nemělo smysl pokračovat, protože spočtené hodnoty mohou s dostatečně velkou pravděpodobností[[1]](#footnote-1) být pouze dílem náhody. Výsledek tohoto testu je, že model s příslušností k subkorpusu statisticky významně lepší je. U výsledného čísla a jemu odpovídající hladině významnosti, zde označené jako *p*, je pouze důležité, aby nebyla překročena hranice 0,05. Čím větší toto číslo je, tím menší je hladina významnosti, což je sice dobře, ale na další interpretaci to nemá vliv.

|  |
| --- |
| Overall model statistics: Likelihood ratio test (LL): 612,83 (p < 0,0001) -> SIGNIFICANT; C-index: 0,61 -> NOT ACCEPTABLE; Nagelkerke R2: 0,04; AIC: 16959,91 |
| Coefficients:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Estimate (log odds) | Standard Error | Z value (Wald) | p-value | Estimate (odds) | 95%CIlower | 95%CI upper |
| (Intercept) | 2,344 | 0,036 | 64,360 | 0,000 | 10,426 | 9,714 | 11,206 |
| Vydané texty - časopisy | 0,543 | 0,146 | 3,725 | 0,000 | 1,721 | 1,306 | 2,316 |
| Vydané texty - noviny | 0,842 | 0,266 | 3,166 | 0,002 | 2,321 | 1,430 | 4,086 |
| Knihy v knihovnách | 0,145 | 0,051 | 2,852 | 0,004 | 1,157 | 1,047 | 1,278 |
| Ostatní | -0,339 | 0,076 | -4,452 | 0,000 | 0,713 | 0,615 | 0,828 |
| Ostatní - blogy | -0,011 | 0,098 | -0,111 | 0,912 | 0,989 | 0,818 | 1,204 |
| Ostatní - bestselery | 2,453 | 0,357 | 6,872 | 0,000 | 11,617 | 6,192 | 25,552 |
| Ostatní - vládní jednání | 0,822 | 0,212 | 3,883 | 0,000 | 2,274 | 1,537 | 3,536 |
| Ostatní - informační bulletiny | 0,804 | 0,325 | 2,474 | 0,013 | 2,235 | 1,248 | 4,522 |
| Ostatní - učebnice | 2,554 | 0,450 | 5,670 | 0,000 | 12,852 | 5,932 | 36,008 |
| Ostatní - zákoníky | 2,307 | 0,193 | 11,930 | 0,000 | 10,040 | 7,015 | 15,012 |
| Ostatní - vládní zprávy | -0,330 | 0,076 | -4,336 | 0,000 | 0,719 | 0,620 | 0,836 |
| Ostatní - Čiebukuro | 0,124 | 0,467 | 0,265 | 0,791 | 1,132 | 0,500 | 3,248 |

 |

Tabulka 4: Logistická regresní analýza slova uči (Brezina 2018)

Dalším údajem je *C-index*. Tato hodnota udává klasifikační nebo též prediktivní sílu modelu. (Brezina 2018) Součástí dat v druhé části tabulky je šance, že výskyt uči v různých subkorpusech bude psán kanou. *C-index* je pravděpodobnost, že náhodně vybranému záznamu bude na základě této šance přiřazen správný způsob zápisu. (Fawcett, 2006, s. 869)[[2]](#footnote-2) Výchozí hodnotou je 0,5 a perfektní model by měl C-index roven 1. (Fawcett, 2006, s. 869) Z toho plyne, že model uvedený v *Tabulce 4* je sice statisticky významně lepší, než by byl model bez vysvětlující proměnné, ale ne o moc. Za přijatelnou hodnotu Lancaster Stats Online považuje 0,7 a vyšší. (Brezina 2018)

Následuje *Nagalkerkův R kvadrát*, který určuje množství variace, kterou model vysvětluje. (Brezina 2018) Uvedená 4% jsou velmi malá hodnota. Poslední vlastnost, *AIC,* je hodnota, která slouží pro porovnávání dvou různých modelů nad stejnými daty. (Brezina 2018) V tomto výzkumu je použit pouze jeden model, proto je hodnota *AIC* irelevantní.

Po údajích, které popisují vlastnosti modelu, následuje část zvaná koeficienty, která popisuje model jako takový. Jako *odds* je označena šance, že náhodně vybraný výskyt v daném subkorpusu je zapsán kanou. Je důležité oddělit pojmy šance a pravděpodobnost. Rozdíl mezi nimi je ten, že šance nějakého jevu je číslo určující, kolikrát je pravděpodobnost tohoto jevu větší, než pravděpodobnost toho, že tento jev nenastane. (Brezina 2018)

První sloupec obsahuje hodnotu přirozeného logaritmu této šance. Je to technický údaj, který je zde ponechán, protože je součástí výstupu použitého nástroje. Pro interpretaci je lepší odpovídající sloupec Estimate (odds), který obsahuje tuto šanci přímo.

Následuje sloupec se *standardní chybou*, která určuje, jak přesně spočítaná šance reflektuje skutečnost. Čím menší číslo, tím lépe. (Brezina 2018) *Waldovo z* a s ním svázaná *hladina významnosti* určují statistickou významnost každého řádku. Ty řádky, které mají *hladinu významnosti* větší než 0,05, lze z další interpretace vyloučit. Znamená to totiž, že nejsou statisticky významné. Hodnoty v nich pozorované mohou být s velkou pravděpodobností dílem náhody. Neznamená to, že tyto údaje jsou špatné, pouze nelze vyloučit jejich náhodný původ. (Brezina 2018)

Další sloupec Estimate (odds) obsahuje šanci, že výskyt patřící do kategorie určené řádkem, bude napsán kanou. V prvním řákdu, který obsahuje popis základních hodnot vysvětlujících proměnných, (Brezina 2018) se s danou hodnotou počítá jinak, než ve zbytku tabulky. Základní hodnotou je subkorpus *Vydané texty – knihy*. Pro ten je uvedena šance zápisu uči kanou 10,426. Pravděpodobnost se z šance určí vztahem:

Rovnice 1: Výpočet pravděpodobnosti z šance (Řeháková 2000, s. 477)

$$P\left(A\right)=\frac{Š(A)}{1+Š(A)}=\frac{10,426}{11,426}\dot{=}0,9125=91,25\%$$

Kde P(A) značí pravděpodobnost, že nastane nějaký jev, v tomto případě že náhodně vybraný zápis slova uči v subkorpusu *Vydané texty - knihy* je zapsán kanou. Š(A) je šance téhož jevu. Tato vypočtená hodnota odpovídá *Tabulce 2*, kde je pro tento subkorpus uveden podíl zápisu hiraganou 91,10% a katakanou 0,15%, což je v součtu právě 91,25%.

Všechny ostatní řádky se neinterpretují takto přímo, protože jsou vyjádřeny vůči tomu prvnímu. To znamená, že například hodnota uvedená pro subkorpus *Ostatní - vládní zprávy* (10,040), říká, že šance na výskyt kanou v tomto subkorpusu je desetkrát větší, než v *Vydané texty - knihy*. Platí tedy, že: $10,426 ∙ 10,040\dot{=}104,68$, což po dosazení do vzorečku výše dá pravděpodobnost výskytu v kaně v tomto subkorpusu 0,9905 (99,05%). Toto opět odpovídá údaji 99,05% v *Tabulce 2*. Protože je uvedená šance relativní vůči prvnímu řádku, je hodnota 1 specifická tím, že značí situaci, kdy k žádné změně šance nedojde, hodnoty menší značí menší šanci a s většími hodnotami zase šance na zápis kanou vzrůstá. Lze tedy shrnout, že čím větší číslo v tabulce je, tím větší podíl kany danému řádku odpovídá.

Poslední dva sloupce udávají interval spolehlivosti na hladině významnosti 95%. V korpusu nejsou zastoupeny všechny texty, jedná se o jejich výběr. Interval spolehlivosti s hladinou významnosti 95% říká, že sledovaná hodnota se bude ve všech textech na základě výskytů ve vzorku vyskytovat s pravděpodobností 95% uvnitř tohoto intervalu. Interval spolehlivosti je uveden pro šanci, což znamená, že, vezme-li se za příklad opět *Ostatní - vládní zprávy*, je interpretací výsledků konstatování, že na základě pozorování je ve všech vládních zprávách podíl zápisu kanou něco mezi 98,55% a 99,41%. Vynásobené šance z dolní hranice intervalu jsou $9,714∙7,015\dot{=}68,14$, po dosazení do vzorečku v Rovnici 1 se získá pravděpodobnost 98,55% a horní hranice se získá obdobně.

V*Tabulce 4* jsou tedy vlastně zakódovány údaje z *Tabulky 2*, doplněné o další údaje, z nichž nejdůležitější je hladina významnosti a interval spolehlivosti. Tyto údaje lze pro snazší čitelnost zanést do zvláštní tabulky (*Tabulka 5*). V této tabulce jsou statisticky nevýznamné záznamy zvýrazněny červeně. Z této tabulky lze potvrdit závěry vypozorované z *Tabulky 2*. Dolní meze intervalů spolehlivosti trojice subkorpusů, *Ostatní - vládní jednání,* *Ostatní - zákoníky* a *Ostatní - vládní zprávy* jsou větší než horní mez *intervalu spolehlivosti* kteréhokoliv jiného subkorpusu.

|  |  |  |
| --- | --- | --- |
| **Subkorpus** | **Podíl kany** | **Interval spolehlivosti** |
| **dolní mez** | **horní mez** |
| Vydané texty - noviny | 96,03% | 93,28% | 97,86% |
| Vydané texty - knihy | 91,25% | 90,67% | 91,81% |
| Vydané texty - časopisy | 94,72% | 92,69% | 96,29% |
| Knihy v knihovnách | 92,34% | 91,05% | 93,47% |
| Ostatní - blogy | 88,14% | 85,66% | 90,27% |
| Ostatní - bestselery | 91,16% | 88,82% | 93,10% |
| Ostatní - vládní jednání | 99,18% | 98,36% | 99,65% |
| Ostatní - informační bulletiny | 95,95% | 93,72% | 97,54% |
| Ostatní - učebnice | 95,89% | 92,38% | 98,06% |
| Ostatní - zákoníky | 99,26% | 98,29% | 99,75% |
| Ostatní - vládní zprávy | 99,05% | 98,55% | 99,41% |
| Ostatní - Čiebukuro | 88,23% | 85,76% | 90,36% |
| Ostatní - verše | 92,19% | 82,93% | 97,33% |

Tabulka 5: Podíly kany ve vzorku a intervaly spolehlivosti na hladině významnosti 95% (NINJAL 2009, Brezina 2018)

## Souhrn

Na základě provedené analýzy bylo určeno několik faktů, které lze o ortografické variabilitě pomocného podstatného jména uči zkonstatovat.

Uči je gramatikalizovaný zástupce pomocných podstatných jmen s předložkovým užitím. Dominantním způsobem zápisu je hiragana a naopak zápis katakanou je pouze okrajový. Podíl zápisu znakem má celokorpusový průměr je 7,42%. Doplněk do 100% připadá na zápis kanou.

Vliv subkorpusů je statisticky významný. Jinými slovy, mezi jednotlivými subkorpusy jsou statisticky významné rozdíly. Byly určeny následující subkorpusy, které jsou nějakým způsobem odlišné: *Ostatní - blogy* a *Ostatní - Čiebukuro*, které jsou internetovými zdroji a obsahují většinu zápisů katakanou z celého korpusu. Subkorpusy *Ostatní - vládní jednání*, *Ostatní - zákoníky* a *Ostatní - vládní zprávy* obsahují oficiální, případně formální texty, a uči je v nich psáno hiraganou v drtivé většině výskytů. V subkorpusech *Ostatní - zákoníky* a *Ostatní - vládní zprávy* se slovo uči vyskytuje častěji, než je v celém korpusu běžné. Jeho četnost na milion slov je dvounásobkem celokorpusového průměru. Vliv subkorpusů *Ostatní - bestselery* a *Ostatní - verše* není statisticky významný.

1. Nejčastěji se jako hraniční hodnota uvádí 0,05 (Brezina 2018, Anděl 2007, s. 9). Zde je tato pravděpodobnost menší než 0,0001. [↑](#footnote-ref-1)
2. C-index se v citovaném článku označuje jako AUC a je definován v jiném kontextu. Interpretace, že se jedná o pravděpodobnost správné funkce modelu, je však stejná. Informace o tom, že C-index a AUC je totožná věc je v poznámkovém aparátu Brezina, 2008 [↑](#footnote-ref-2)