Poprawianie pisowni

Błędna pisownia może być powodowana przez różne czynniki:

- błędy w przekazywaniu sygnałów do mięśni odpowiedzialnych za ruch palców (literówki, np. *litreówki*);
- nieznajomość pisowni słów i zasad ortografii (wymowa błędnej formy jest taka sama lub podobna do wymowy formy poprawnej, np. *Stary człowiek i może* jako tytuł znanego utworu literackiego a nie reklama leku);
- nieznajomość morfologii (np. *upartość* zamiast *upór*).

Metody poprawiania pisowni powinny uwzględniać model błędów użytkownika.

Odległość Levenshteina

Odległość Levenshteina dwóch łańcuchów znaków x i y to minimalna liczba prostych operacji edycyjnych, które przekształcają łańcuch x w łańcuch y (lub odwrotnie).

Proste operacje edycyjne to:

- wstawienie znaku, np. kota $\rightarrow$ kwota
- usunięcie znaku, np. kwota $\rightarrow$ kota
- zamiana znaku, np. kota $\rightarrow$ koza
- zamiana miejscami dwóch sąsiadujących znaków (tzw. czeski błąd), np. pisk $\rightarrow$ psik

Odległość Levenshteina

$$
\begin{array}{rcll}
ed(X_{1\ldots i+1}, Y_{1\ldots j+1}) & = & ed(X_{1\ldots i}, Y_{1\ldots j}) & \text{jeśli } x_{i+1} = y_{j+1} \\
 & & & \text{(ostatnie znaki} \\
 & & & \text{takie same)} \\
 & = & 1 + min\{ed(X_{1\ldots i-1}, Y_{1\ldots j-1}, & \text{jeśli } x_i = y_{j+1} \\
 & & \quad ed(X_{1\ldots i+1}, Y_{1\ldots j}), & \text{i } x_{i+1} = y_j \\
 & & \quad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} & \text{(dwa ostatnie znaki} \\
 & & & \text{zamienione miejscami)} \\
 & = & 1 + min\{ed(X_{1\ldots i}, Y_{1\ldots j}), & \text{w pozost. przypadkach} \\
 & & \quad ed(X_{1\ldots i+1}, Y_{1\ldots j}), & \\
 & & \quad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} & \\
ed(X_{1\ldots 0}, Y_{1\ldots j}) & = & j & 0 \leq j \leq n \\
ed(X_{1\ldots i}, Y_{1\ldots 0}) & = & i & 0 \leq i \leq m \\
ed(X_{1\ldots -1}, Y_{1\ldots j}) & = & ed(X_{1\ldots i}, Y_{1\ldots -1}) = max(m, n) & \text{(Definicje brzegowe)}
\end{array}
$$

Odległość Levenshteina

przestawienie: (*kula*,*kual*)

$$
\begin{aligned}
ed(X_{1\ldots i+1}, Y_{1\ldots j+1}) &= ed(X_{1\ldots i}, Y_{1\ldots j}) && \text{jeśli } x_{i+1} = y_{j+1} \text{ (ostatnie znaki takie same)} \\
&= 1 + min\{ed(X_{1\ldots i-1}, Y_{1\ldots j-1}, && \text{jeśli } x_i = y_{j+1} \\
&\qquad ed(X_{1\ldots i+1}, Y_{1\ldots j}), && \text{i } x_{i+1} = y_j \\
&\qquad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} && \text{(dwa ostatnie znaki zamienione miejscami)} \\
&= 1 + min\{ed(X_{1\ldots i}, Y_{1\ldots j}), && \text{w pozost. przypadkach} \\
&\qquad ed(X_{1\ldots i+1}, Y_{1\ldots j}), \\
&\qquad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} \\
ed(X_{1\ldots 0}, Y_{1\ldots j}) &= j && 0 \leq j \leq n \\
ed(X_{1\ldots i}, Y_{1\ldots 0}) &= i && 0 \leq i \leq m \\
ed(X_{1\ldots -1}, Y_{1\ldots j}) &= ed(X_{1\ldots i}, Y_{1\ldots -1}) = max(m, n) && \text{(Definicje brzegowe)}
\end{aligned}
$$

Odległość Levenshteina

usunięcie: (*bab*,*baba*)

$$
\begin{aligned}
ed(X_{1\ldots i+1}, Y_{1\ldots j+1}) &= ed(X_{1\ldots i}, Y_{1\ldots j}) && \text{jeśli } x_{i+1} = y_{j+1} \\
& && \text{(ostatnie znaki} \\
& && \text{takie same)} \\
&= 1 + min\{ed(X_{1\ldots i-1}, Y_{1\ldots j-1}, && \text{jeśli } x_i = y_{j+1} \\
& \qquad ed(X_{1\ldots i+1}, Y_{1\ldots j}), && \text{i } x_{i+1} = y_j \\
& \qquad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} && \text{(dwa ostatnie znaki} \\
& && \text{zamienione miejscami)} \\
&= 1 + min\{ed(X_{1\ldots i}, Y_{1\ldots j}), && \text{w pozost. przypadkach} \\
& \qquad ed(X_{1\ldots i+1}, Y_{1\ldots j}), \\
& \qquad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} \\
ed(X_{1\ldots 0}, Y_{1\ldots j}) &= j && 0 \leq j \leq n \\
ed(X_{1\ldots i}, Y_{1\ldots 0}) &= i && 0 \leq i \leq m \\
ed(X_{1\ldots -1}, Y_{1\ldots j}) &= ed(X_{1\ldots i}, Y_{1\ldots -1}) = max(m, n) && \text{(Definicje brzegowe)}
\end{aligned}
$$

Odległość Levenshteina

$$
\begin{aligned}
ed(X_{1\ldots i+1}, Y_{1\ldots j+1}) &= ed(X_{1\ldots i}, Y_{1\ldots j}) && \text{wstawienie: } (babab, baba) \\
& && \text{(ostatnie znaki takie same)} \\
&= 1 + min\{ed(X_{1\ldots i-1}, Y_{1\ldots j-1}, && \text{jeśli } x_i = y_{j+1} \\
& \quad ed(X_{1\ldots i+1}, Y_{1\ldots j}), && \text{i } x_{i+1} = y_j \\
& \quad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} && \text{(dwa ostatnie znaki zamienione miejscami)} \\
&= 1 + min\{ed(X_{1\ldots i}, Y_{1\ldots j}), && \text{w pozost. przypadkach} \\
& \quad ed(X_{1\ldots i+1}, Y_{1\ldots j}), \\
& \quad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} \\
ed(X_{1\ldots 0}, Y_{1\ldots j}) &= j && 0 \leq j \leq n \\
ed(X_{1\ldots i}, Y_{1\ldots 0}) &= i && 0 \leq i \leq m \\
ed(X_{1\ldots -1}, Y_{1\ldots j}) &= ed(X_{1\ldots i}, Y_{1\ldots -1}) = max(m, n) && \text{(Definicje brzegowe)}
\end{aligned}
$$

Odległość Levenshteina

$$
\begin{aligned}
ed(X_{1\ldots i+1}, Y_{1\ldots j+1}) &= ed(X_{1\ldots i}, Y_{1\ldots j}) && \text{jeśli } x_{i+1} = y_{j+1} \\
& && \text{(ostatnie znaki} \\
& && \text{zamiana: } (kod,kot) \\
&= 1 + min\{ed(X_{1\ldots i-1}, Y_{1\ldots j-1}, && \text{jeśli } x_i = y_{j+1} \\
& \quad ed(X_{1\ldots i+1}, Y_{1\ldots j}), && \text{i } x_{i+1} = y_j \\
& \quad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} && \text{(dwa ostatnie znaki} \\
& && \text{zamienione miejscami)} \\
&= 1 + min\{ed(X_{1\ldots i}, Y_{1\ldots j}), && \text{w pozost. przypadkach} \\
& \quad ed(X_{1\ldots i+1}, Y_{1\ldots j}), \\
& \quad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} \\
ed(X_{1\ldots 0}, Y_{1\ldots j}) &= j && 0 \leq j \leq n \\
ed(X_{1\ldots i}, Y_{1\ldots 0}) &= i && 0 \leq i \leq m \\
ed(X_{1\ldots -1}, Y_{1\ldots j}) &= ed(X_{1\ldots i}, Y_{1\ldots -1}) = max(m, n) && \text{(Definicje brzegowe)}
\end{aligned}
$$

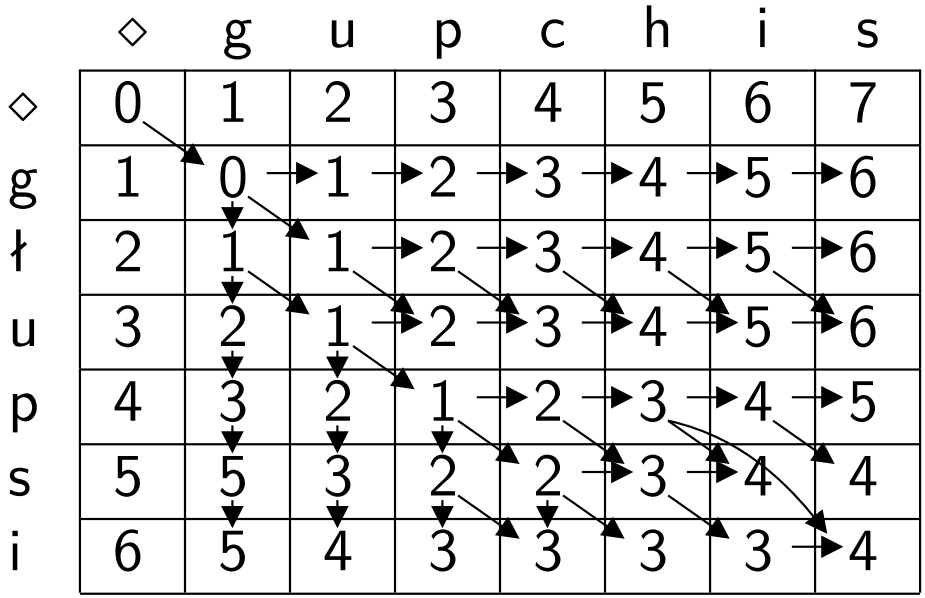
Odległość Levenshteina

$$
\begin{aligned}
ed(X_{1\ldots i+1}, Y_{1\ldots j+1}) &= ed(X_{1\ldots i}, Y_{1\ldots j}) && \text{jeśli } x_{i+1} = y_{j+1} \text{ (ostatnie znaki takie same)} \\
&= 1 + min\{ed(X_{1\ldots i-1}, Y_{1\ldots j-1}, && \text{usunięcie: } (kot, kota) \\
&\qquad ed(X_{1\ldots i+1}, Y_{1\ldots j}), && \text{i } x_{i+1} = y_j \\
&\qquad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} && \text{(dwa ostatnie znaki zamienione miejscami)} \\
&= 1 + min\{ed(X_{1\ldots i}, Y_{1\ldots j}), && \text{w pozost. przypadkach} \\
&\qquad ed(X_{1\ldots i+1}, Y_{1\ldots j}), \\
&\qquad ed(X_{1\ldots i}, Y_{1\ldots j+1})\} \\
ed(X_{1\ldots 0}, Y_{1\ldots j}) &= j && 0 \leq j \leq n \\
ed(X_{1\ldots i}, Y_{1\ldots 0}) &= i && 0 \leq i \leq m \\
ed(X_{1\ldots -1}, Y_{1\ldots j}) &= ed(X_{1\ldots i}, Y_{1\ldots -1}) = max(m, n) && \text{(Definicje brzegowe)}
\end{aligned}
$$

Odległość Levenshteina

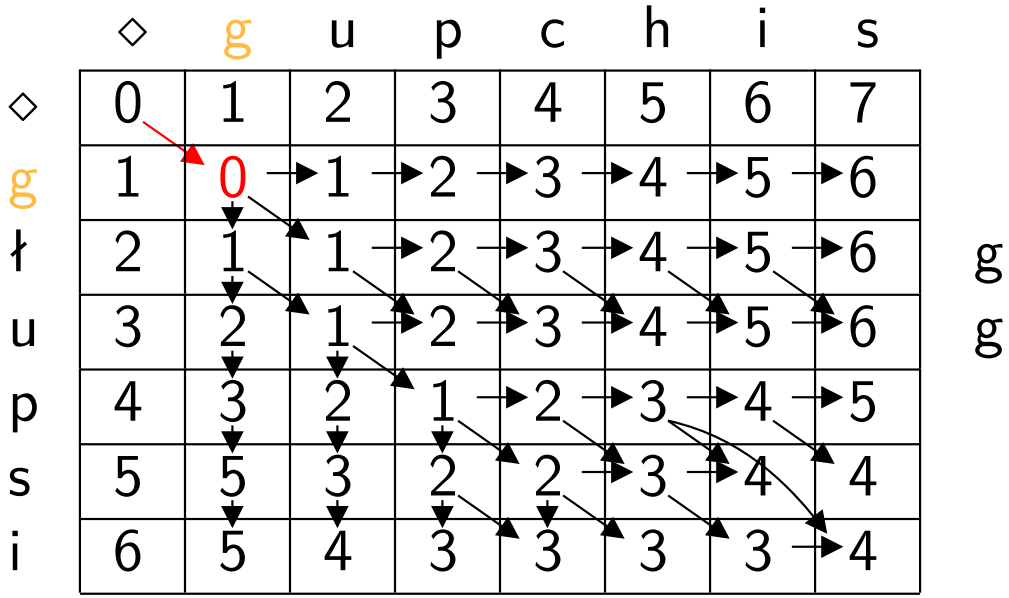
$$
\begin{array}{rcll}
ed(X_{1\ldots i+1}, Y_{1\ldots j+1}) & = & ed(X_{1\ldots i}, Y_{1\ldots j}) & \text{jeśli } x_{i+1} = y_{j+1} \\
 & & & \text{(ostatnie znaki} \\
 & & & \text{takie same)} \\
 & = & 1 + min\{ed(X_{1\ldots i-1}, Y_{1\ldots j-1}, & \text{jeśli } x_i = y_{i+1} \\
 & & ed(X_{1\ldots i+1}, Y_{1\ldots j}), & \text{wstawienie: } (kota,kot) \\
 & & ed(X_{1\ldots i}, Y_{1\ldots j+1})\} & \text{(dwa ostatnie znaki} \\
 & & & \text{zamienione miejscami)} \\
 & = & 1 + min\{ed(X_{1\ldots i}, Y_{1\ldots j}), & \text{w pozost. przypadkach} \\
 & & ed(X_{1\ldots i+1}, Y_{1\ldots j}), & \\
 & & ed(X_{1\ldots i}, Y_{1\ldots j+1})\} & \\
ed(X_{1\ldots 0}, Y_{1\ldots j}) & = & j & 0 \leq j \leq n \\
ed(X_{1\ldots i}, Y_{1\ldots 0}) & = & i & 0 \leq i \leq m \\
ed(X_{1\ldots -1}, Y_{1\ldots j}) & = & ed(X_{1\ldots i}, Y_{1\ldots -1}) = max(m, n) & \text{(Definicje brzegowe)}
\end{array}
$$

Odległość edycyjna – przykład



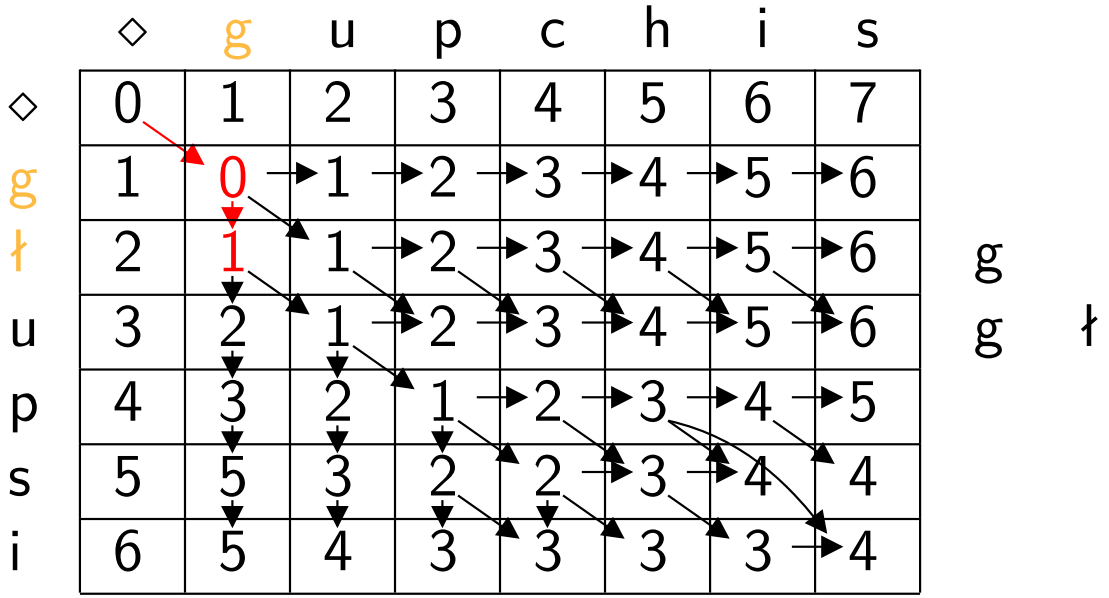
Odległość edycyjna – przykład

$$\begin{aligned} ed(\mathrm{g},\mathrm{g}) &= ed(\epsilon,\epsilon) \\ &= 0 \end{aligned}$$



Odległość edycyjna – przykład

$$
\begin{aligned}
ed(\text{g}, \text{gł}) &= 1 + min\{ed(\epsilon, \text{g}), ed(\text{g}, \text{g}), ed(\epsilon, \text{gł})\} \\
&= 1 + min\{1, 0, 2\} = 1
\end{aligned}
$$



Odległość edycyjna – przykład

$$
\begin{aligned}
ed(\text{gu}, \text{głu}) &= ed(\text{g}, \text{gł}) \\
&= 1
\end{aligned}
$$

	◇	g	u	p	c	h	i	s
◇	0	1	2	3	4	5	6	7
g	1	0	1	2	3	4	5	6
ł	2	1	1	2	3	4	5	6
u	3	2	1	2	3	4	5	6
p	4	3	2	1	2	3	4	5
s	5	5	3	2	2	3	4	4
i	6	5	4	3	3	3	3	4

g u

g ł u

Odległość edycyjna – przykład

$$\begin{aligned} ed(\text{gup}, \text{głup}) &= ed(\text{gu}, \text{głu}) \\ &= 1 \end{aligned}$$

	◇	g	u	p	c	h	i	s
◇	0	1	2	3	4	5	6	7
g	1	0	1	2	3	4	5	6
ł	2	1	1	2	3	4	5	6
u	3	2	1	2	3	4	5	6
p	4	3	2	1	2	3	4	5
s	5	5	3	2	2	3	4	4
i	6	5	4	3	3	3	3	4

g u p

g ł u p

Odległość edycyjna – przykład

$$
\begin{aligned}
ed(\text{gupc}, \text{głups}) &= 1 + min\{ed(\text{gup}, \text{głup}), ed(\text{gupc}, \text{głup}), \\
&\quad ed(\text{gup}, \text{głups})\} \\
&= 1 + min\{1, 2, 2\} = 2
\end{aligned}
$$

	◇	g	u	p	c	h	i	s
◇	0	1	2	3	4	5	6	7
g	1	0	1	2	3	4	5	6
ł	2	1	1	2	3	4	5	6
u	3	2	1	2	3	4	5	6
p	4	3	2	1	2	3	4	5
s	5	5	3	2	2	3	4	4
i	6	5	4	3	3	3	3	4

g u p c

g ł u p s

Odległość edycyjna – przykład

$$
\begin{aligned}
ed(\text{gupch}, \text{głups}) &= 1 + min\{ed(\text{gupc}, \text{głup}), ed(\text{gupch}, \text{głup}), \\
&\quad ed(\text{gupc}, \text{głups})\} \\
&= 1 + min\{2, 3, 2\} = 3
\end{aligned}
$$

	◇	g	u	p	c	h	i	s
◇	0	1	2	3	4	5	6	7
g	1	0	1	2	3	4	5	6
ł	2	1	1	2	3	4	5	6
u	3	2	1	2	3	4	5	6
p	4	3	2	1	2	3	4	5
s	5	5	3	2	2	3	4	4
i	6	5	4	3	3	3	3	4

g		u	p	c	h
g	ł	u	p	s	

Odległość edycyjna – przykład

$$\begin{aligned} ed(\text{gupchi}, \text{głupsi}) &= ed(\text{gupch}, \text{głups}) \\ &= 3 \end{aligned}$$

	◇	g	u	p	c	h	i	s
◇	0	1	2	3	4	5	6	7
g	1	0	1	2	3	4	5	6
ł	2	1	1	2	3	4	5	6
u	3	2	1	2	3	4	5	6
p	4	3	2	1	2	3	4	5
s	5	5	3	2	2	3	4	4
i	6	5	4	3	3	3	3	4

g		u	p	c	h	i
g	ł	u	p	s		i

Odległość edycyjna – przykład

$$
\begin{aligned}
ed(\text{gupchis}, \text{głupsi}) &= 1 + min\{ed(\text{gupch}, \text{głup}), \\
&\quad ed(\text{gupchi}, \text{głupsi}), ed(\text{gupchis}, \text{głups})\} \\
&= 1 + min\{3, 3, 4\} = 4
\end{aligned}
$$

	◇	g	u	p	c	h	i	s
◇	0	1	2	3	4	5	6	7
g	1	0	1	2	3	4	5	6
ł	2	1	1	2	3	4	5	6
u	3	2	1	2	3	4	5	6
p	4	3	2	1	2	3	4	5
s	5	5	3	2	2	3	4	4
i	6	5	4	3	3	3	3	4

g		u	p	c	h	i	s
g	ł	u	p	s		i	

Odległość edycyjna – przykład

	◇	g	u	p	c	h	i	s
◇	0	1	2	3	4	5	6	7
g	1	0	1	2	3	4	5	6
ł	2	1	1	2	3	4	5	6
u	3	2	1	2	3	4	5	6
p	4	3	2	1	2	3	4	5
s	5	5	3	2	2	3	4	4
i	6	5	4	3	3	3	3	4

g		u	p	c	h	i	s
g	ł	u	p		s	i	

Odległość edycyjna – przykład

	◇	g	u	p	c	h	i	s
◇	0	1	2	3	4	5	6	7
g	1	0	1	2	3	4	5	6
ł	2	1	1	2	3	4	5	6
u	3	2	1	2	3	4	5	6
p	4	3	2	1	2	3	4	5
s	5	5	3	2	2	3	4	4
i	6	5	4	3	3	3	3	4

g		u	p	c	h	i	s
g	ł	u	p			s	i

Edycyjna odległość odcięcia

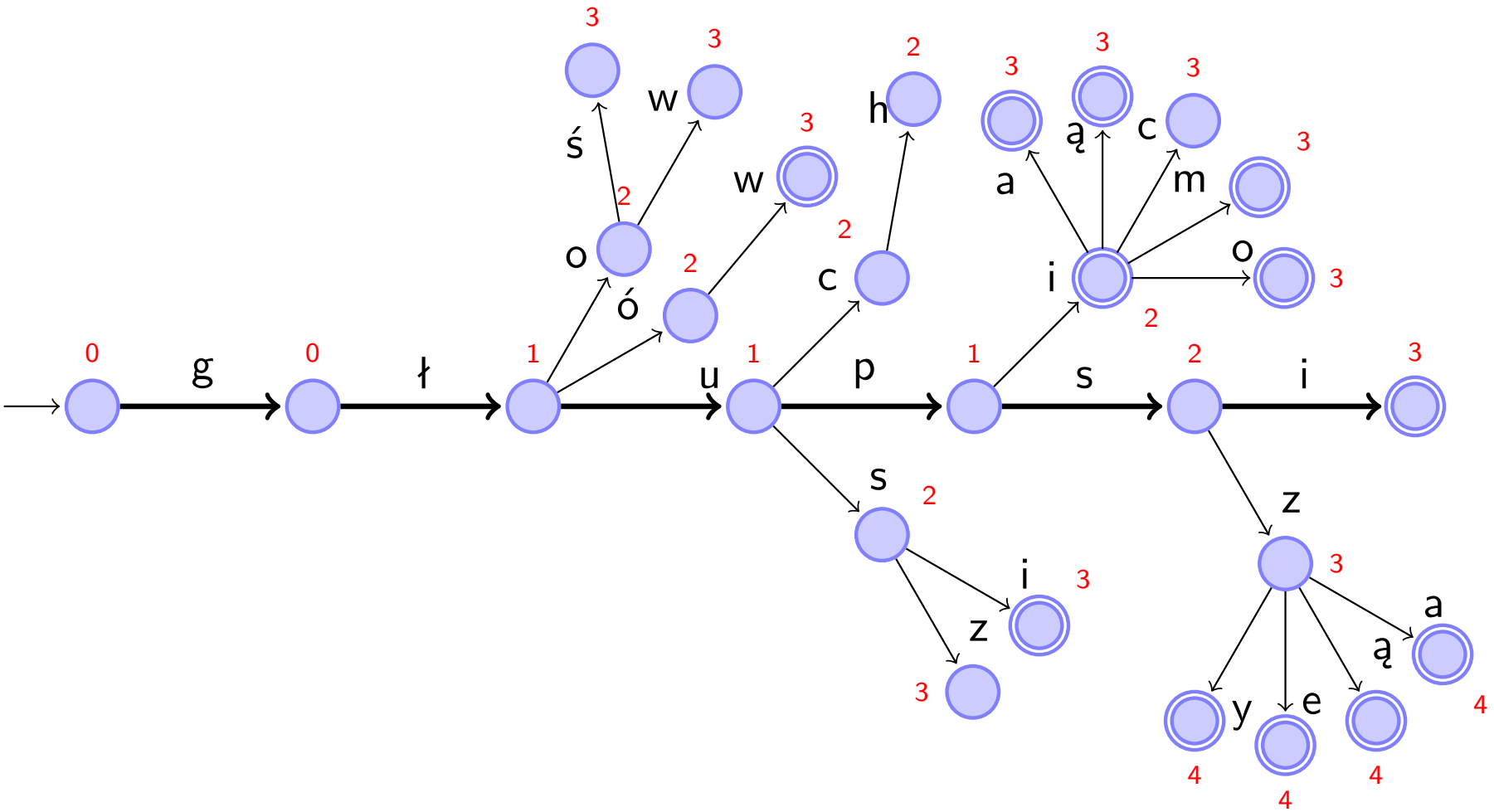
Porównywanie w całości wszystkich słów w słowniku z formą niepoprawną X (o długości m) byłoby zbyt kosztowne. Wystarczy sprawdzić tylko część potencjalnego zamiennika Y (o długości n) z częścią niepoprawnej formy X i zadecydować, czy sprawdzać następne litery formy:

$$cuted(X_{1\ldots m}, Y_{1\ldots n}) = \min_{l \leq i \leq u} ed(X_{1\ldots i}, Y_{1\ldots n}),$$

$$l = min(1, n - t), u = max(m, n + t)$$

t jest maksymalną dopuszczalną odległością, operacje *min* i *max* zapobiegają wyjściu indeksów poza granice słowa.

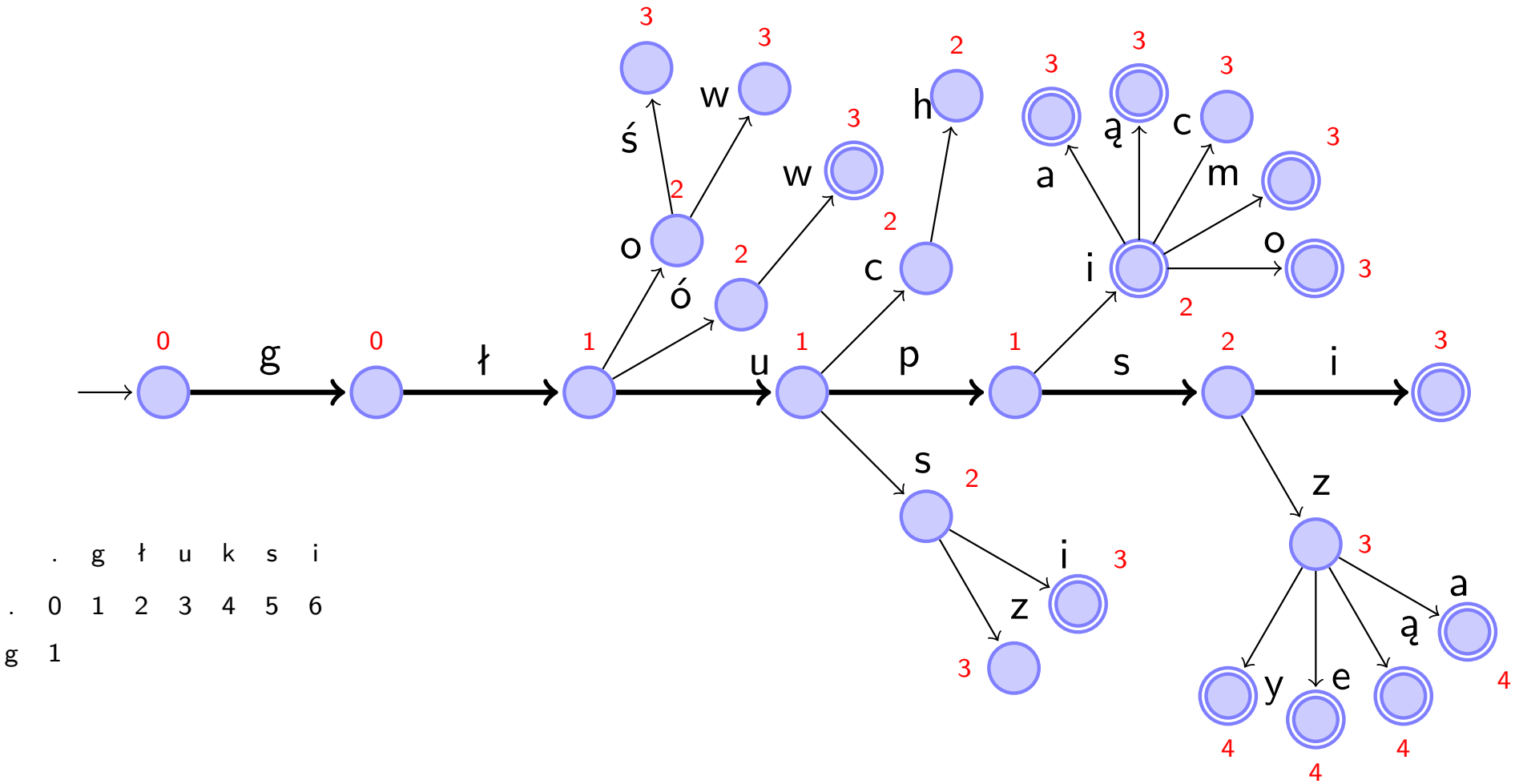
Edycyjna odległość odcięcia



Dla $t = 4$,

cuted(*gupchis,głuch*)=min{ed(g,głuch),ed(gu,głuch),ed(gup,głuch),
ed(gupc,głuch),ed(gupch,głuch),ed(gupchi,głuch),ed(gupchis,głuch)}=2

Edycyjna odległość odcięcia



Dla $t = 4$,
cuted(*gupchis,głuch*)=min{ed(g,głuch),ed(gu,głuch),ed(gup,głuch),
ed(gupc,głuch),ed(gupch,głuch),ed(gupchi,głuch),ed(gupchis,głuch)}=2