Kwartalnik Językoznawczy 2013/1

Daniel Śledziński

Normalizacja tekstów w języku polskim – aspekty lingwistyczne

1. Wprowadzenie

Normalizacja tekstów to proces, który obejmuje określone transformacje w obrębie tekstu. Podstawowe zadanie normalizacji to zmiana pisanej formy tekstu na formę mówioną. Normalizacja tekstów ma kilka istotnych zastosowań związanych z technologią języka i mowy.

W syntezie mowy zapis wyrazów, fraz i zdań musi być doprowadzony do takiej postaci, w jakiej będzie on odczytany przez syntezator, zatem wszystkie elementy muszą być przekonwertowane do postaci słownej (np. liczebniki zapisane przy użyciu cyfr). Trzeba zaznaczyć, że dla syntezy mowy normalizacja musi być prowadzona w czasie rzeczywistym, zatem w sposób w pełni automatyczny.

Normalizacja tekstów wykonywana jest również na potrzeby systemów automatycznego rozpoznawania mowy (ang. ASR) – jednak proces ten dotyczy etapu budowy tych systemów. Normalizuje się teksty, które są potrzebne dla budowy modeli językowych – dopiero te modele uczestniczą we właściwym procesie rozpoznawania mowy. Na podstawie normalizowanych korpusów tekstowych mogą być też wykonywane badania m.in. lingwistyczne. Inne zastosowanie normalizacji tekstów dotyczy systemów tłumaczenia automatycznego.

W zależności od konkretnego zastosowania oraz od przyjętych założeń proces normalizacji może obejmować następujące czynności:

- konwersja wyrażeń numerycznych i wyrażeń słowno-numerycznych do postaci słownej;
- rozwinięcie skrótów i skrótowców;
- zamiana wielkości znaków;
- usunięcie wybranych lub wszystkich znaków interpunkcyjnych, typograficznych i diakrytycznych;
- usunięcie wybranych elementów niezwiązanych z tekstem, np. tabel, rysunków, podpisów;
- wstawienie znaczników dla specyficznych elementów – np. dla wyrazów obcych.

W artykule najwięcej uwagi poświęcono konwersji wyrażeń numerycznych i wyrażeń słowno-numerycznych na zapis słowny. Skoncentrowano się głównie na aspektach lingwistycznych oraz potencjalnych problemach wynikających z budowy poszczególnych wyrażeń. Wydaje się, że opracowanie ważnych z punktu widzenia procesu normalizacji zagadnień lingwistycznych może być przydatne na różnych etapach budowy systemu normalizacji automatycznej. W tekście w bardzo ograniczonym zakresie odniesiono się do aspektów technicznych – omówiono m.in. działanie algorytmu konwersji liczb całkowitych na zapis słowny. Przedstawiono też kilka elementarnych wyrażeń regularnych wykrywających w tekstach elementy wymagające konwersji na zapis słowny.

2. Podstawowe problemy

W pełni automatyczna (i bezbłędna) normalizacja tekstów w języku polskim jest trudna do uzyskania, ponieważ mamy do czynienia z językiem o złożonym systemie fleksyjnym[1]. Wśród odmiennych części mowy znajdują się elementy, które podlegają normalizacji – wyrazy, które muszą być przekonwertowane do postaci słownej. Podstawowa trudność związana jest z tym, że najczęściej nie sposób określić formy fleksyjnej danego elementu tekstu

[1] Por.: J. Strutyński, *Gramatyka polska. Wprowadzenie, fonetyka, fonologia, morfologia, składnia*, Kraków 2002; Z. Saloni, *Czasownik polski. Odmiana, słownik*, wyd. 3 zm., Warszawa 2007; E. Tabakowska, *Kognitywne podstawy języka*, Kraków 2001.

tylko na podstawie zapisu liczbowego. Problem ten dotyczy liczebników, dat kalendarzowych i czasu zapisanych przy użyciu cyfr. Dotyczy on również skrótów i skrótowców, które muszą być rozwinięte do postaci słownej. Omawiany problem związany jest z różnorodnością uwarunkowań konotacyjnych czasownika i innych części mowy. W tabeli 1 jako przykład przytoczono warianty łączliwości fleksyjnej czasownika *komunikować*[2].

Tabela 1. Łączliwość fleksyjna czasownika *komunikować*

pytanie	**przyimek**	**przypadek**	**przykład**
co?	–	A	Po raz drugi dyrekcja komunikuje swoje stanowisko […]
komu?	–	D	Komunikujemy wszystkim mieszkańcom bloku, że […]
o czym?	o	L	Radio komunikowało o przebiegu akcji ratowniczej.
przez co?	przez	A	Władze miasta komunikowały przez megafony o zbliżającym się niebezpieczeństwie.

Konotacja jest rozumiana jako otwieranie przez poszczególne wyrazy możliwości występowania określonych wyrazów w określonej formie gramatycznej. Dla określenia formy gramatycznej elementów podlegających normalizacji potrzebna jest bezbłędna analiza składniowa zdania.

Drugi istotny problem dotyczący normalizacji tekstów związany jest z faktem, że niektóre elementy (homografy) mogą być różnie odczytywane – nawet gdy znana jest ich forma gramatyczna. Zapis *123.159* może oznaczać zarówno liczbę rzeczywistą złożoną z części całkowitej i mantysy, jak i liczbę całkowitą rozdzieloną separa-

[2] S. Mędak, *Praktyczny słownik łączliwości składniowej czasowników polskich. Słownik języka polskiego*, Kraków 2005.

torem, a zapis *800100000* może być zarówno liczbą całkowitą, jak i numerem telefonu. Poza tym istnieje wiele ciągów znaków, które mogą być różnie odczytywane przy zachowaniu tego samego lub zbliżonego znaczenia – dotyczy to np. dat kalendarzowych lub określeń czasu. Istnieje również kilka sposobów odczytywania liczb rzeczywistych. Z kolei skrótowce mogą być czytane zarówno w postaci skróconej, jak i rozwiniętej. Tego typu problemy mogą być rozwiązane poprzez wykonanie stosownych badań dotyczących częstotliwości odczytywania poszczególnych wariantów wyrażeń podlegających procesowi normalizacji.

Kolejny istotny problem związany jest z tym, że niektóre elementy w tekście mogą być zapisywane w sposób niestandardowy, a czasami nawet niepoprawny językowo lub przy udziale inwencji twórczej autora. Zatem mogą wystąpić trudności z automatyczną identyfikacją takich elementów w tekście i – co z tego wynika – ich poprawną konwersją na zapis słowny.

3. Normalizacja wybranych elementów tekstu

3.1. Liczebniki główne

W języku polskim liczebniki główne odmieniają się przez rodzaje i przypadki (np. *dwieście trzydzieści pięć domów* – mianownik rodzaju męskiego; *dwustu trzydziestu pięciu domów* – dopełniacz rodzaju męskiego, *dwustu trzydziestu pięciu mężczyzn* – mianownik i dopełniacz rodzaju męskoosobowego)[3].

Generowanie form słownych dla liczebników głównych odbywa się na podstawie prostego algorytmu. Liczebnik analizowany jest niejako od końca – trójki cyfr pobierane są od końca, ale nie zmienia się kolejności w obrębie tych trójek. Liczba *120517065123* będzie przetworzona w sposób ujęty w tabeli 2.

[3] S. Mędak, *Liczebnik też się liczy*, Kraków 2004.

Tabela 2. Zapis odwrotny liczby *120517065123*

zakres:	<1000	tysięcy	milionów	miliardów	bilionów
liczba:	123	(0)65	517	120	–

Dysponując takim zapisem, można wygenerować słowny zapis dla liczb trzycyfrowych znajdujących się w poszczególnych zakresach. Następnie wystarczy dodać nazwę zakresu i całość połączyć (ponownie w odwróconej kolejności). Generowanie liczb trzycyfrowych w poszczególnych zakresach (ewentualnie dwucyfrowych lub jednocyfrowych – jeżeli określają one tylko jedności lub jedności i dziesiątki) można oprzeć na tabelach konwersji zawierających zapisy słowne dla setek, dziesiątek oraz jedności w odpowiednim przypadku i dla odpowiedniego rodzaju. Tabela 3 zawiera przykład konwersji dla rodzaju męskiego w dopełniaczu.

Tabela 3. Przykładowa tabela konwersji

	zakres jedności	**zakres dziesiątek**	**zakres setek**
0	zera	–	–
1	jednego	dziesięciu	stu
2	dwóch	dwudziestu	dwustu
3	trzech	trzydziestu	trzystu
4	czterech	czterdziestu	czterystu
5	pięciu	pięćdziesięciu	pięciuset
6	sześciu	sześćdziesięciu	sześciuset
7	siedmiu	siedemdziesięciu	siedmiuset
8	ośmiu	osiemdziesięciu	ośmiuset
9	dziewięciu	dziewięćdziesięciu	dziewięciuset

Przy generowaniu zapisu słownego trójek cyfr uwzględnia się zapis słowny kolejno: setek, dziesiątek oraz jedności. Następnie trzeba dodać nazwę odpowiedniego zakresu (tysięcy, milionów itd.) w odpowiednim przypadku. Tabela 4 przedstawia zapis wygenerowanych trójek cyfr dla poszczególnych zakresów (dot. liczby *120517065123*).

Tabela 4. Wynik przykładowej konwersji w obrębie trójek cyfr

zakres	**<1000**	**tysięcy**	**milionów**	**miliardów**	**bilionów**
liczba	123	(0)65	517	120	–
zapis słowny	stu dwudziestu trzech	sześć-dziesięciu pięciu tysięcy	pięciuset siedem-nastu milionów	stu dwudziestu miliardów	–

Ostatni krok konwersji to połączenie zapisów słownych wszystkich liczb trzycyfrowych, łącznie z nazwami zakresów w odpowiednich przypadkach (zaczynając od zakresu największego). Zatem ostateczny zapis słowny liczby *120517065123* w dopełniaczu i w rodzaju męskim to: *stu dwudziestu miliardów pięciuset siedemnastu milionów sześćdziesięciu pięciu tysięcy stu dwudziestu trzech*.

Jeżeli cała liczba występuje w mianowniku, to przypadek nazwy konkretnego zakresu (tysięcy, milionów itd.) zależy od liczebności tego zakresu. Wynika to z zasad łączenia liczebników z rzeczownikami – przypadek rzeczownika jest uzależniony od liczebności (liczebniki: tysiąc, milion, miliard, bilion odmieniają się jak rzeczowniki):

- dla liczby jeden jest to mianownik liczby pojedynczej (np. *jeden milion*),
- dla liczb 2, 3 lub 4 oraz liczb zakończonych cyfrą 2, 3 lub 4 jest to mianownik liczby mnogiej (np. *2 tysiące, 224 miliony*),
- dla liczby 0 i liczb z zakresu 5–19, a także liczb co najmniej dwucyfrowych zakończonych cyfrą 0, 1, 5, 6, 7, 8

lub 9 jest to dopełniacz liczby mnogiej (np. *121 milionów*, *11 tysięcy*, *20 tysięcy*, *551 bilionów*, *125 tysięcy*, *150 bilionów*, *159 tysięcy*).

Jeżeli konwertowana liczba występuje w innym przypadku niż mianownik, to nazwy zakresów występują również w tym przypadku – niezależnie od wartości ostatnich cyfr w poszczególnych grupach (np. *dwustu dwudziestu trzech tysięcy*, *o dwustu dwudziestu czterech tysiącach*, *o dwustu dwudziestu pięciu tysiącach*).

Na koniec warto wspomnieć, że przy większych liczbach całkowitych czasami używane są separatory zwiększające czytelność zapisu. Poniższe wyrażenie regularne wykrywa liczby całkowite z dowolną liczbą separatorów:

```
^([1-9][0-9][0-9](\.[0-9][0-9][0-9])+)$
```

3.2. Liczebniki porządkowe

Liczebniki porządkowe mogą być zapisane:

- w postaci słownej (np. *pierwszy*, *dwudziesty piąty*);
- przy użyciu cyfr arabskich (np. *1998 rok*);
- przy użyciu cyfr rzymskich;
- przy użyciu cyfry połączonej z końcówką fleksyjną (np. *2-gi*, *5-tego*) lub z kropką (np. *6.* – ‘szósty’).

Z punktu widzenia automatyzacji procesu normalizacji tekstu problemem może być identyfikacja liczebników porządkowych w tekście, w szczególności liczebników porządkowych zapisanych przy użyciu cyfr arabskich – to, czy dana liczba jest liczebnikiem porządkowym, może wynikać tylko z kontekstu.

W przypadku liczebników porządkowych większych niż 100 jako formy porządkowe odczytuje się tylko dwie ostatnie cyfry (opisujące dziesiątki i jedności). Cyfry dotyczące setek oraz tysięcy odczytuje się tak jak liczebniki główne (np. *trzysta czterdziesty piąty*, *tysiąc dziewięćset dziewięćdziesiąty drugi*, *dwa tysiące dwunasty* itp.). Dlatego dla konwersji dowolnych liczebników porządkowych (z zapisu cyfrowego na zapis słowny) wystarczy dysponować tabli-

cami konwersji dla cyfr oznaczających jedności i dziesiątki – takie tablice muszą zawierać zapis słowny liczebników porządkowych dla różnych przypadków i rodzajów. Dla setek oraz dla tysięcy można użyć tablic konwersji utworzonych dla liczebników głównych. Nie dotyczy to przypadków, w których większe liczebniki porządkowe wyrażają tylko krotność setek lub tysięcy (np. ‘setny’, ‘pięćsetny’, ‘tysięczny’ itp.).

Liczby rzymskie można w tekście wykryć za pomocą wyrażenia regularnego:

```
^([IVXLCDM]+)$
```

Liczby rzymskie wymagają konwersji do zapisu wyrażonego cyframi arabskimi. Otrzymane w ten sposób liczebniki porządkowe można przekształcić na zapis słowny według omówionych zasad.

3.3. Liczby rzeczywiste

Liczby rzeczywiste złożone są z części całkowitej oraz z części ułamkowej (mantysy) oddzielonej kropką lub przecinkiem. Oto wyrażenie regularne wykrywające liczby rzeczywiste:

```
^([0-9]+(\.|,)[0-9]+)$
```

Zapis słowny części całkowitej wyznacza się według zasad opisanych w podrozdziale 3.1. Przy normalizacji trzeba uwzględnić fakt, że część ułamkowa może być różnie odczytywana – np. liczbę *4.25* można odczytać w sposób następujący:

cztery i dwadzieścia pięć setnych

cztery przecinek dwadzieścia pięć

cztery i jedna czwarta

Prawdopodobieństwo poszczególnych możliwości odczytywania liczb rzeczywistych może być sprawdzone w oddzielnych badaniach.

3.4. Liczebniki w tekstach

Poniżej przytoczono kilkadziesiąt losowych liczebników pobranych z artykułów publikowanych w serwisie internetowym Onet w 2013 roku:

40 lat W. Brytanii w UE
67 proc. głosujących
wyniósł 35 proc.
około 60 osób
na drodze nr 786
drogą krajową nr 74
na 40 km odcinku
na ponad 20 km odcinek
wynosząca 34 wypadki
z łącznie 773 ofiarami
16 wypadków
wyniósł on 68 dni
10 ludzi na ziemi
300 mln euro
dzieci do lat 12
u około 6 proc. chorych
do 780 milionów litrów
zdobywając 185 mandatów
101 i 40 miejsc
zdobyła 37 mandatów
z 32 mandatami
43 posłów niezależnych
wyniesie 2,4 proc
z przyrostem 5,5 proc
wynosi ok. 25,4 proc.
44 miejscowości
według 84 proc.
że 9 proc. z nich
1,4 mld USD
sprzed 10 lat
spadek o 7 pkt. proc.
do 77 proc.
33-letni
89 proc. deklaruje
18-minutowe
wynosił 29 proc.
56 proc. twierdzi
z 41 do 32 proc.
około 600 osób
65 proc. sądzi
przez 10 lat
0,75 proc. całkowitej kwoty
od 2,5 do 3 tys.
65-letnia
blisko 20 lat temu
60 nowych firm
32 mln zł
z 2,5 tys. miejsc pracy
5 tys. miejsc pracy
10 uzdrowisk
44 gminy
w wieku 35–54 lat
33-latek

Powyższe zestawienie ukazuje realne problemy związane z automatyzacją normalizacji liczebników w tekstach języka polskiego.

Niektóre liczebniki muszą być odczytane w przypadku innym niż mianownik. Inne są zapisane częściowo słownie – przy użyciu skrótów, które w rozwiniętej formie podlegają odmianie (np. *mln, mld*). Pewne wyzwanie dla procesu automatyzacji mogą stanowić pary liczebników oznaczające zakres. Osobną kategorię stanowią liczebniki połączone z rzeczownikiem lub przymiotnikiem (*33--latek, 33-letni*). Powyższe zestawienie z pewnością nie wyczerpuje wszystkich możliwości użycia liczebników.

3.5. Daty kalendarzowe

Wyróżnia się dwa sposoby zapisywania dat. Pierwszy sposób związany jest z użyciem nazw miesięcy – daty utworzone w ten sposób to daty kalendarzowe. Drugi sposób polega na użyciu numeru tygodnia lub numeru dnia w roku – są to daty porządkowe. Zdecydowanie częściej używa się dat kalendarzowych. Daty porządkowe wykorzystywane są głównie w przemyśle, a w tekstach i w mowie potocznej są one praktycznie niespotykane, dlatego nie uwzględniono ich w dalszych rozważaniach.

Wytyczne, których celem jest ujednolicenie formatu dat, zostały sprecyzowane w międzynarodowej normie ISO 8601[4]. Według tej normy:

- rok powinien być zapisywany czterema cyframi;
- dni i miesiące powinny być dwucyfrowe (np. *01, 05, 12, 31*);
- składniki daty są zapisywane kolejno – od składnika najbardziej znaczącego do składnika najmniej znaczącego (rok, miesiąc, dzień);
- składniki powinny być oddzielone znakami rozdzielającymi.

Daty zgodne z tymi wytycznymi (np. *1999-09-09, 1999.09.09, 1999/09/09*) są rzadko spotykane w tekstach w języku polskim. Da-

[4] Norma ustalona przez międzynarodową Organizację Normalizacyjną (ISO – ang. International Organization for Standardization).

ty zapisane w ten sposób można wykryć przy użyciu wyrażenia regularnego:

```
^([1-9]\d\d\d[-/.](0[1-9]|1[012]|[1-9])[-/.](0[1-|[12]\d|3[01]|[1-9]))$
```

Większość dat kalendarzowych w tekstach w języku polskim nie spełnia przytoczonej normy. W języku polskim składniki daty zapisuje się najczęściej w kolejności od składnika najmniej znaczącego do składnika najbardziej znaczącego: dzień, miesiąc, rok. Poza tym daty kalendarzowe mogą zawierać wszystkie składniki (informacje o dniu, miesiącu i roku) lub tylko wybrane składniki. Poszczególne składniki daty mogą, ale nie muszą być zapisywane w sposób jednorodny – często zapis cyfr arabskich połączony jest z zapisem słownym lub zapisem cyfr rzymskich. Z tych informacji wynika, że w tekstach w języku polskim spotyka się wiele sposobów formatowania dat kalendarzowych. Fakt ten potwierdza poniższy wykaz – zawiera on kilkadziesiąt losowo wybranych dat z artykułów publikowanych w internetowym serwisie Onet w 2013 roku:

w 1973 roku
z 1975 r.
w październiku 2012 r.
w 2007 roku
z 1989 roku
w 2008 r.
w 2016 r.
od 1945 roku
rok 2012 był
w 1994 roku
od roku 1997
30 stycznia
od 28 lipca 2010 roku
3 czerwca
11 grudnia 2012 r.
10 stycznia br.
w czerwcu 2011 r.
z października 2011 r.
na przełomie czerwca i lipca 2010 r.
w 1997 r.
w czerwcu 2013 roku
w latach 2011–2013
1 sierpnia 2012 r.
23 kwietnia
w 1984 roku
6–12 czerwca
w roku 2003
w 2013 r.
w 2010 r.
w listopadzie 2012 roku
stan na 30 lipca 2013
w 2011 roku
27 kwietnia 2014 r.
w latach 1957–89
w 2015 r.
w 1941 roku
w 1906 roku
mniej osób niż 2003 r.
w 2013 r.

Powyższy wykaz ukazuje różne sposoby formatowania dat kalendarzowych w tekstach w języku polskim. Bardzo często występuje tylko jeden składnik daty – informacja o roku. Równie często zapis nie jest jednorodny (miesiąc jest zapisany słownie). Oddzielną kategorię stanowią pary dat wyznaczające okres. Warto zauważyć, że wśród losowo wybranych dat nie wystąpił żaden przypadek daty zawierającej wszystkie składniki i zapisanej w sposób jednorodny.

Zadanie normalizacji wymaga nie tylko właściwej identyfikacji składników dat, ale także poprawnego ich przekształcenia na zapis słowny. Poniższa lista zawiera wskazówki dotyczące poprawnego odczytywania dat w języku polskim:

- składniki daty odczytuje się w kolejności od składnika najmniej znaczącego do składnika najbardziej znaczącego (kolejno: dzień, miesiąc, rok);
- składnik dnia jest liczebnikiem porządkowym z zakresu 1–31, którego przypadek jest uzależniony od kontekstu;
- jeżeli data zawiera informację o dniu miesiąca, to nazwę miesiąca podaje się w dopełniaczu (np. *dziesiąty maja*, *dwudziesty piąty czerwca*, *dwudziestego piątego czerwca*), w przeciwnym razie przypadek nazwy miesiąca zależy od kontekstu;
- zapis słowny dla składnika roku tworzony jest zgodnie z zasadami dotyczącymi generowania zapisu słownego liczebników porządkowych (zob. rozdz. 3.2). Jeżeli oprócz składnika roku występuje składnik miesiąca, to zapis słowny roku odczytywany jest w dopełniaczu, natomiast jeżeli występuje tylko składnik roku, to jego przypadek jest uzależniony od kontekstu;
- jeżeli składnik roku występuje samodzielnie, to przypadek wyrazu *rok* jest uzależniony od kontekstu. W przeciwnym razie wyraz *rok* odczytywany jest w dopełniaczu (np. *przed rokiem dwa tysiące dwunastym*, ale: *przed czerwcem dwa tysiące dwunastego roku*).

Oto przykłady zdań zawierających daty utworzone zgodnie z podanymi wytycznymi (wyraz *dzień* jest podmiotem ukrytym):

Dzisiaj jest dwudziesty piąty (dzień) czerwca tysiąc dziewięćset dziewięćdziesiątego ósmego roku.

Mówił o wydarzeniach z dwudziestego piątego (dnia) czerwca tysiąc dziewięćset dziewięćdziesiątego ósmego roku.

Ukończył szkołę w roku dwa tysiące dwunastym.

Ukończył szkołę w czerwcu dwa tysiące dwunastego roku.

W praktyce – w mowie potocznej i przy odczytywaniu dat – obserwuje się liczne odstępstwa od podanych wskazówek dotyczących odczytywania dat. Niektóre odstępstwa wynikają z faktu, że poprawne odczytywanie pełnych dat może być uciążliwe – dotyczy to w szczególności tych tekstów, w których występuje ich wiele. Dlatego często stosowane są pewne skróty myślowe. Inne odstępstwo polega na odczytaniu (lub wypowiedzeniu) składnika daty w niewłaściwym przypadku. Ewentualne uwzględnienie tego czynnika w systemie normalizacji tekstów uzależnione jest od celów tego systemu. Wydaje się, że przy normalizacji tekstów wykonywanej na potrzeby budowy modeli językowych dla systemów ASR może być to czynnik istotny. Poniżej przedstawiono kilka przykładów słownego zapisu daty niezgodnego z podanymi wcześniej wskazówkami:

dwudziesty piąty czerwiec dwa tysiące trzynastego roku

dwudziesty piąty czerwca dwa tysiące trzynaście

dwudziestego piątego czerwca dziewięćdziesiątego ósmego

w czerwcu dwa tysiące trzynastego

czerwiec dwa tysiące trzynaście

dziewięćdziesiąty ósmy

dwudziestego piątego

w dwa tysiące trzynastym

3.6. Czas (dnia)

W podrozdziale zawarto informacje dotyczące czasu rozumianego jako czas dnia (godzina). Dla normalizacji tekstów istotne są następujące fakty związane z zapisywaniem czasu:

- czas może być zapisany przy użyciu dwóch składników (godzin i minut), rzadko przy użyciu trzech składników (godzin, minut i sekund). W wyjątkowych sytuacjach podawane są setne części sekundy lub nawet mniejsze jednostki. Często podaje się informacje tylko o godzinie (szczególnie w przypadku godzin pełnych);
- składnik godziny może być zapisany w systemie 12--godzinnym lub w systemie 24-godzinnym;
- zapis godzin może być jednocyfrowy lub dwucyfrowy (np. *6:00*, *23:00*), niekiedy zapis godziny jest dwucyfrowy, pomimo że godzina jest liczbą mniejszą od 10 (np. *06:20*);
- zapis minut, sekund i setnych części sekund jest zawsze dwucyfrowy (np. *10:07*);
- poszczególne składniki czasu oddziela się znakami rozdzielającymi – dwukropkiem, kropką, rzadziej kreską;
- czas dnia często zapisuje się słownie – szczególnie w sytuacjach, w których jest to godzina pełna (*godzina trzecia*, *o szesnastej*) lub jeżeli jest to godzina *12:00* lub *24:00* (*w południe*, *o północy*). Spotykane są również wyrażenia słowno-numeryczne.

Informacje dotyczące zapisywania czasu zostały sprecyzowane w normie ISO 8601:

- musi być to zapis w systemie 24-godzinnym;
- poszczególne składniki muszą być dwucyfrowe;
- poszczególne składniki muszą być rozdzielone znakami rozdzielającymi.

Dotychczasowe rozważania dotyczyły sposobów zapisu czasu w tekście. Z rozważań tych wynika, że większość zapisów czasu ma zbliżoną postać i jeżeli nie są to zapisy słowne, mogą być wykryte przez jedno wyrażenie regularne:

```
^(0|0\d|\d|1\d|2[0-4])[-|:|\.](0\d|[1-5]\d)$
```

W procesie normalizacji tekstu po wykryciu zapisanego w tekście czasu dnia następuje identyfikacja poszczególnych składników oraz generowanie możliwych zapisów słownych. W praktyce czas dnia może być odczytany różnorako, jednak nie wszystkie sposoby używane w mowie potocznej są poprawne językowo. Podobnie jak przy podawaniu dat, często stosowane są skróty myślowe lub wyrazy zastępcze (np. *wpół do*, *kwadrans po*). Poza tym przy wymawianiu godziny w systemie dwunastogodzinnym nierzadko dodaje się dodatkowy wyraz określający porę dnia (*rano*, *po południu*, *wieczorem*, *w nocy*). Poniżej przedstawiono listę spotykanych sposobów określania czasu dnia:

godzina czternasta trzydzieści pięć

dziesiąta rano

szesnasta trzydzieści

wpół do szóstej

za dwadzieścia trzecia

pięć po wpół do czwartej

druga w nocy

dwunasta

południe

3.7. Skróty i skrótowce

Pojęcie skrótu często jest mylone ze skrótowcem. Skrótowiec (akronim) zbudowany jest najczęściej z pierwszych liter lub sylab wyrażenia złożonego z dwóch lub więcej wyrazów. Skrótowiec

może być odczytany zarówno formie skróconej, jak i pierwotnej – rozwiniętej. Oto popularny podział skrótowców[5]:

- skrótowce literowe (literowce) – są złożone z pierwszych liter wyrazów wyrażenia złożonego (np. *NBP – Narodowy Bank Polski*, *AWF – Akademia Wychowania Fizycznego*). Poszczególne litery w literowcach są wymawiane osobno. Na potrzeby normalizacji tekstu można przyjąć, że literowce będą konwertowane do zapisu zgodnego z wymową (*en be pe*, *pe zet u*) lub nie będą modyfikowane. Mogą one też zostać rozwinięte na podstawie informacji przechowywanej w bazie danych;
- skrótowce głoskowe (głoskowce) – również są zbudowane z pierwszych liter wyrażenia złożonego, jednak – odczytywane łącznie, np. *ZUS* (*Zakład Ubezpieczeń Społecznych*), *GUS* (*Główny Urząd Statystyczny*). Zatem normalizacja głoskowców nie jest potrzebna lub ewentualnie można przyjąć, że ich zapis będzie zmieniany na małe litery;
- skrótowce grupowe (grupowce lub sylabowce) – składają się z kilku pierwszych głosek – najczęściej z pierwszych sylab wyrazów wyrażenia złożonego (np. *Polfa – Polska Farmacja*). Sylabowce również wymawiane są jako suma połączonych głosek (lub sylab), dlatego ich normalizacja także nie jest potrzebna;
- skrótowce mieszane – są to dowolne kombinacje literowców, głoskowców i sylabowców, np. *PZMot – Polski Związek Motorowy*. Ponieważ skrótowce mieszane składają się m.in. z literowców, ich normalizacja może wiązać się z konwersją do zapisu z wydzielonymi literami czytanymi osobno, np. *pe zet mot*;
- skrótowce złożeniowe – są zbudowane z całego wyrazu określanego oraz części wyrazu określającego, np. *Investbank* (‘bank inwestycyjny’). Wymowa takich skrótowców wynika wprost z zapisu ortograficznego, zatem nie muszą być one konwertowane.

[5] *Słownik poprawnej polszczyzny PWN*, red. W. Doroszewski, wyd. 18, Warszawa 1995; *Słownik ortograficzny języka polskiego wraz z zasadami pisowni i interpunkcji*, red. M. Szymczak, wyd. 10, Warszawa 1989.

Cechą charakterystyczną skrótowców jest to, że mogą być one odczytywane w formie skróconej bądź jako pełne wyrażenie, z którego uzyskano skrótowiec, np. zapis *student AWF* może być odczytany jako: *student a wu ef* lub *student Akademii Wychowania Fizycznego*. Problem różnych możliwości odczytywania elementów podlegających normalizacji był już sygnalizowany w poprzednich rozdziałach.

Dalsze rozważania dotyczą skrótów. Podobnie jak skrótowce także skróty są skróconym zapisem określonych wyrażeń. W przeciwieństwie do skrótowców skróty często uzyskuje się z pojedynczych wyrazów. Zasadnicza różnica między skrótami a skrótowcami polega na tym, że skrótów nie odczytuje się w formie skróconej, lecz jako pierwotne wyrazy lub wyrażenia, z których je uzyskano. Zatem skrót jest skróconą formą zapisu wyrazu lub wyrażenia, jednak sposób odczytywania nie zmienia się. Poza tym skróty prawie zawsze zapisuje przy użyciu małych liter, choć istnieją nieliczne odstępstwa od tej reguły (np. skrót *Sz.P.* jest zapisany wielkimi literami ze względu na formę grzecznościową).

W Wikipedii omówiono podział skrótów ze względu na umiejscowienie kropki (lub kropek). Według tego opracowania skróty w języku polskim można podzielić na:

- skróty pisane małymi literami z kropką na końcu – jeżeli skrót jest utworzony z pojedynczego wyrazu i nie zawiera ostatniej litery tego wyrazu (np. *inż.*, *tel.*). Taki zapis dotyczy również skrótów powstałych z pierwszych liter kilku wyrazów pod warunkiem, że żaden z tych wyrazów (oprócz pierwszego) nie rozpoczyna się samogłoską (np. *cdn.*, *itd.*, *itp.*, *jw.*);
- skróty pisane małymi literami bez kropki na końcu – te skróty zawierają ostatnią literę wyrazu skracanego (np. *dr*, *mgr*, *nr*). Również skróty będące symbolami matematycznymi lub chemicznymi, międzynarodowymi oznaczeniami miar oraz oznaczeniem rodzimej waluty są pisane bez kropki na końcu, pomimo że nie kończą się ostatnią literą wyrazu skracanego (np. *zł*, *sin*, *log*, *cm*);
- kropki umieszczone są wewnątrz skrótów powstałych z kilku wyrazów, jeżeli którykolwiek z tych wyrazów (poza

pierwszym) zaczyna się samogłoską (np. *p.n.e.*). Dotyczy to również skrótów zawierających więcej niż jedną literę któregoś ze skracanych wyrazów (np. *m.in.*, *m.st.*). Poza tym nieliczne skróty zawierają wewnętrzne kropki w celu ich odróżnienia od skrótów identycznych (np. *b.r. – brak roku [wydania]*, *br. – bieżącego roku*).

Automatyzacja procesu normalizacji wymaga utworzenia bazy danych skrótów i skrótowców. Każdy rekord w takiej bazie musi zawierać zestaw łańcuchów tekstowych przedstawiających dopuszczalne sposoby zapisu poszczególnych skrótów i skrótowców. Z przedstawionych w tym rozdziale informacji wynika, że każdy skrót i skrótowiec ma tylko jedną poprawną formę zapisu. Wnikliwa analiza tekstów pokazuje jednak, że zdarzają się pewne modyfikacje – skrót może być zapisany wielkimi literami (kapitalikami) – zwłaszcza gdy tekst umieszczony obok skrótu również jest zapisany w ten sposób (np. *MGR JAN KOWALSKI*). Zdarza się również, że skrót zapisany jest ze spacjami (np. *c. d. n.*). Czasami spotyka się również niepoprawny zapis skrótowców – małymi literami (np. *pzu*). Wymienione zniekształcenia podlegają pewnym schematom – a zatem można w sposób automatyczny wygenerować dla każdego skrótu i skrótowca zbiór potencjalnych form niepoprawnych. Poza tym popularne skróty podlegają regułom fleksji (np. *dra – doktora*). Te skróty wskazują na pewną formę gramatyczną skracanego wyrazu, dlatego właściwe wydaje się umieszczenie tych form w osobnych rekordach bazy danych.

Dla każdego skrótu i skrótowca (formy poprawnej i zestawu form alternatywnych) baza danych musi przechowywać informację o możliwych sposobach odczytania. Dla skrótowców powinny to być zarówno formy skrócone, jak i formy rozwinięte. Skróty odczytywane są jako wyrazy lub wyrażenia, z których te skróty uzyskano. Należy wziąć pod uwagę fakt, że rozwinięte formy skrótowców odmieniają się przez przypadki i najczęściej z zapisu skróconego nie wynika, który przypadek został użyty (np. *NBP – Narodowy Bank Polski* lub: *spotkał się z prezesem NBP – prezesem Narodowego Banku Polskiego*).

3.8 Inne wyrażenia

W tekstach występuje wiele wyrażeń, które noszą charakterystyczne cechy i w procesie normalizacji wymagają konwersji do zapisu słownego. Są to m.in.:

- adresy stron internetowych;
- adresy poczty elektronicznej;
- adresy miejsc zamieszkania (i instytucji);
- numery telefoniczne;
- numery identyfikacyjne;
- kwoty pieniędzy wyrażone określoną walutą;
- wartości wyrażone w różnych miarach (długości, objętości i wagi).

Ten wykaz z pewnością nie wyczerpuje wszystkich możliwości. W dalszym ciągu rozdziału omówiono potencjalne problemy związane z normalizacją tych elementów.

Adresy stron internetowych oraz adresy poczty elektronicznej zbudowane są według określonych reguł, dlatego stosunkowo łatwo można je wykryć – m.in. na podstawie znaku @ lub na podstawie nazwy domeny najwyższego rzędu (np. *pl*, *org*, *com*). Poza tym może istnieć kilka możliwości odczytania danego adresu (np. z uwzględnieniem bądź bez uwzględnienia wyrazu *kropka*).

Adresy zamieszkania składają się z kilku standardowych elementów. Co prawda układ tych elementów bywa zróżnicowany, jednak zadanie normalizacji można rozpatrywać oddzielnie dla każdego z nich.

Automatyczna identyfikacja numerów telefonicznych może być utrudniona – zdarza się, że są one mylone z dużymi liczbami całkowitymi. Poza tym istnieje kilka możliwości zapisywania oraz odczytywania numerów telefonicznych. Jako duże liczby całkowite mogą być również błędnie interpretowane różne numery identyfikacyjne złożone z cyfr (np. numer Regon lub NIP). Z drugiej jednak strony często w sąsiedztwie wymienionych elementów znajduje się łańcuch tekstowy wskazujący na ich znaczenie (np. *tel.*).

Fakt, że podany zapis przedstawia kwotę, może być automatycznie wykryty na podstawie obecności łańcucha tekstowego *zł*[6]. Poza tym można utworzyć zestaw wyrażeń regularnych dla różnych sposobów zapisywania kwot.

Jednostki miar wyrażają wielkości fizyczne (wagę, długość, objętość, odcinki czasu). W opracowaniach naukowych spotyka się wiele nazw jednostek miar, z których większość nie jest używana na co dzień. Można je podzielić na:

- jednostki podstawowe, np. *metr*, *sekunda*, *kilogram*;
- jednostki wtórne (krotne), np. *kilometr*, *minuta*, *tona*;
- jednostki pochodne – jednostki utworzone na bazie jednostek podstawowych, np. *m/s* (*metr na sekundę*).

Dla normalizacji tekstów istotne jest to, że podstawowe i wtórne jednostki miary zapisywane są w postaci skrótów. Skróty te muszą być przekonwertowane do pełnego zapisu słownego. Często przy użyciu skrótu zapisany jest inny przypadek niż mianownik (por. rozdz. 3.7).

Jednostki pochodne (np. *m/s*, *km/godz.*, *km/s* itd.) są kombinacjami jednostek podstawowych, dlatego jest ich więcej. Przy odmianie przez przypadki jednostki pochodnej odmienia się tylko pierwszy jej człon (*kilometrów na godzinę*, *metrów na sekundę*). Trzeba pamiętać, że przypadek jednostki miary w mianowniku zależy od liczebności, np. *dwa kilometry*, *pięć kilometrów*, *dwadzieścia cztery kilometry* (por. rozdz. 3.1). Poza tym w tekstach często występują kombinacje wyrazów, które są zbudowane według schematu tworzenia jednostek pochodnych i podlegają tym samym regułom związanym z odmianą, jednak pierwszy lub oba człony nie są typowymi jednostkami fizycznymi, np. *str./min.*, *zł/kg*, *zł/mb*, *zł/szt*. Takie jednostki również podlegają odmianie (np. *strony na minutę*, *stron na minutę*).

[6] W niektórych tekstach kwoty podawane są z międzynarodowym kodem waluty (dla złotówki jest to PLN). Informacje o międzynarodowych kodach walut zawarte są w dokumencie ISO 4217.

4. Podsumowanie

W artykule przedstawiono zagadnienia związane z normalizacją tekstów w języku polskim. Omówiono problemy związane z konwersją: liczb całkowitych, liczb rzeczywistych, dat kalendarzowych, czasu dnia, skrótów i skrótowców oraz innych elementów spotykanych w tekstach. Artykuł z pewnością nie wyczerpuje tematu, który jest bardzo rozległy, jednak uwzględnione informacje zostały przedstawione w sposób usystematyzowany i odnoszący się do zagadnień elementarnych. Przedstawione zagadnienia lingwistyczne są szczególnie istotne z punktu widzenia zadania automatyzacji procesu normalizacji tekstów języka polskiego.

Z informacji zamieszczonych w artykule wynika, że automatyzacja pozornie prostej czynności – odczytywania wyrażeń numerycznych – jest w rzeczywistości zadaniem złożonym i wieloaspektowym, w szczególności dla języka polskiego – języka o bardzo rozbudowanym systemem fleksyjnym.

Normalization of texts in Polish language – linguistic aspects

SUMMARY

The paper presents linguistic aspects of text normalization for Polish. Text normalization is a process which converts the orthographic form of the text into the spoken form. Generally the following elements of texts are converted in process of normalization: numbers, time and dates, abbreviations, acronyms and other expressions that differ in written and spoken form. Normalization of a Polish is complex process because of the inflected character of language. Particular forms of words results from context, therefore syntactic and semantic analysis has to be carried out for each sentence in order to obtain a correct spoken form. Moreover often there are several ways of normalization of a particular expression – also informal ways. The article describes norms of spoken forms of time and dates expressions and algorithm for numbers conversion.

Key words: text normalization, text analysis, speech synthesis, speech recognition, Polish.

O Autorze:

Daniel Śledziński – doktor nauk humanistycznych w Instytucie Językoznawstwa Uniwersytetu im. Adama Mickiewicza w Poznaniu. Zainteresowania: przetwarzanie i analiza tekstów oraz sygnału mowy, percepcja mowy, fonetyka akustyczna, fonologia, języki programowania, tworzenie aplikacji, bazy danych, sztuczne sieci neuronowe, statystyka.
E-mail: danielsl@poczta.onet.pl