**APPENDIX A. Mohmil word reduction rules**

|  |  |  |
| --- | --- | --- |
| Input  Word | Criteria | Stem  Word |
| پانی وانی | Second word start wit و[wao]  and rest of the letters are alike in both words. | پانی |
| اکیلا دکیلا | Second word start with و [wao]  and rest of the letters are alike in both words. | اکیلا |
| چوری چکاری | First letters and last two letters  match in both words | چوری |
| ٹھیک ٹھاک | First two letters and last letters  match in both words | ٹھیک |
| حیص بیص | Second word start with ب[bay]  and rest of the letters are alike in both words. | حیص |
| دھو دھا | First two letters match in both  words | دھو |
| چاوچوز | First one letter match in both  words | چاو |
| خالی خولی | First one and last two letter  match in both words | خالی |
| پیس پاس | First one and last one letter  match in both words | پیس |
| سمجھا بجھا | Last three letters match in both  words | سمجھ |
| پکڑ دھکڑ | Last two letters match | پکڑ |
| جھاڑو بہارو | Last one letter match in both  words | جھاڑ |
| دھوم دھام | First two and last one letter  match in both words | دھوم |
| پھینک پھانک | First and last two letters match  in both words | پھینک |
| چوڑا چکلا | First and last one letters match  in both words | چوڑا |
| بچا کھچا | Last two letters match in both  words | بچا |
| بات چیت | Last one letters match in both  words | بات |

**APPENDIX B. Urdu prefixes list**

|  |
| --- |
| م ب ن |
| ذو بو لے مد گد کف کڑ کج ظل طب صف ذی تل تپ پٹ جل مے تو دم شہ غم دل بں نر کم نا زی بے با اک یک ہر نو لا سر تہ پر ہم  آب اب ال آس شپ آل بر بڑ ھر ھم چو سر چی سش سہ ڈی پس بہ آؤ آن پچ پل پن تت تر تہ حق دل نم دو لب گل گز رگ رہ تر زر زن کٹ فن غل سگ شب شش شہ |
| ادھ سرد خار ہیڈ ماہ کڑی کچے کچی عطر عدت ضرب زیر زہر روح دشت برقی چوب جوش جام ٹکٹ تخت تال پیر پسر پاے پان جلا بخت ابر آہن آبی زود بال تنگ اشک نیک برگ بدھ باٹ باب اگن اگر اسم ابن غیر ہمہ ہفت ہشت زور نیم راہ آئر آبا ابو آبی ابی نور خوب اہل لوح لخت گلا گرم گرد کوے کوہ کور کند کشش آہن کرۂ کرہ ایک باج باد بار عون باگ بام عضو عرش عدم ببر بحث ظفر طبع ضعف بحر صلح صدر شوم شوخ شمع سیف سطح ستم سبک بدر ژرف زود برج زرہ رشک دلی دقت برق بڑا بزم داغ حکم حسن حسب حرف حبل بغل  بلا بنی جنم جنت جمع جشن تنگ ثمر ثقل زبر تیغ تنگ تند تلا ترش تخم تحت تاو تاؤ تام گلو تار تاج عدل پیک نیل پون پنگ ناز پنج پکی پکا  فور پست پدم قبل خرد پاس شپس جگر جگت تیز تیر تخم مسدس خود خوب تلخ |
| مدبر عددی طبعی صحیح شکست سبزہ زہرہ زنان دماغ حکمت حرکی حاصل چوبی جذبہ ترکی ترکش تحیت تازی تادم پنجہ پلنگ پستہ پتھر پائے پاؤں تخمی آبلہ سوڈو روئے ایکس انڈر آہنی باسی بازی بادہ بندہ برنج برتن آثار اپلا ابجد ابدی آبرو ابلا ابلغ ابلق اپیل مدار نظام خلاف ہوال دارل  ناقص نازک مرکب ماہر ماتم لیزر لیڈی کورا مسلـک مردم مذاق محفل مجلس مثبت مادر لباس گوشہ گوشہ گورا گندہ گلشن گرمی گرما گردش کورٹ کثیر کامل کاتب قومی آتشی فوجی عضوی پرست عالی طلاق الہی ضعیف انشا بابا باقی شوخی بانگ بایو بخشی شعلہ شعلہ سیرت سجدہ سجدۂ بدری سبزۂ بدعت سادہ برہم دامن زریں رویت روشن روبہ رونق رائے ذہنی بستر بسیط بعید حالت حاضر بقاے بگڑا بلند چہار  چائے جوہر جذبہ ترنج جائے ثقلی تیغہ تیغہ تیزی تکیہ تکیہ تکلف تشنہ تشبہ ترنم صاحب بلیک بزرگ تراش تخمی تختہ تختہ برزخ تبسم تباہ تائے  تانب باڈی تازہ تارک بادۂ تابہ تابع پیلی پہلا پھٹا گراں پنجہ پنبہ پنبہ بالا پارہ پختہ طالع پتلی پتخہ پبلـک ابیض موجب پاکٹ پارہ پارۂ جلدی جلال جلاء تیرہ |
| مجازی متوسط گرمیء گردشی صحیفہ زنانہ جوابی ٹھنڈی تکبیر تشنہء تشبیہ تسخیر تحریف تحریر تائید تاریخ پچھلے ابالی ابجدی ابدان ابروے ابناء ابنای ابوال ابدال آتشیں اتمام آلودہ انگشت بازار باران بادہء آئینہ آوارہ بادیہ باراں باریک باطنی بدرجہ بندہء برہان برائے ولدال میثاق آخرال اخذال نورال ہومیو وقرال نوائے نظامت متبرک ماتمی لیڈیز کھوٹا کھوٹی مستقل مرکزی مذہبی مخملی مخلوط محکمہ محکمہ مجلسی متفقہ مادری ماتمی لشکری مابعد ٹرائی کشادہ قانون طر یقہ طالع مافوق صحیفہ شیریں شومیء شوخئی شوخیء شہادت سیاسی سلطان زنانی زاو یہ دماغی دعائے  دارال خلافت حیثیت حاضری جوہری بدرجہ ثانوی جرائم ثلاثی ٹیڑھی تقصیر ٹھنڈا تیز یہ تیزاب تکمیل تکلیف تیرگی تقویم تقطیر جلالت تعویذ تعدیل تصفیہ تصفیہ پارٹی پارچہ پائین ترکیب تذکرہ تحلیل تحسین تحریم تحریک پچھلی پہاڑی شکستہ تجنیس |
| بہترین احقرال مائکرو ماہانہ مردانہ مجموعہ متوازی متنازع کورانہ طوفانی طبقاتی سلطانی سرکاری سالانہ زنجـیری رومانی روحانی روایتی روحانی  روایتی ڈسچارج خوشامد تیزابی تکمیلی تکبیری تقابلی تعدیلی تصدیقی تحو یلی پیمانہ ابنائے اتفاقی بازاری آئینہء پاکیزہ پبلشنگ پیمانہ تاریخی تاریک  اولوال تبدیلی نافذال واحدال گلوگیر نادرال ناقابل ناجائز تدریجی ترجیحی مزاحیہ مخملیں مجموعی تقاضاے تقطیری مجموعہ تکوینی تولیدی متعلقہ گیسوئے ٹکسالی کاتبین قانونی عمرانی زاو یہء طر یقہء زہر یلی صوبائی شیرینی شاہانہ زہر یلے ژولیدہ سائنسی معروضی |
| قرارداد اظفارال کیمیائی کیمیاوی طلسماتی خاندانی ترازوئے ابتدائے ابتدائی سائنٹفک شیرینیء طلسمانی گورنمنٹ مجرمانہ محافظتی محکمانہ مدافعتی مشرکانہ مدافعتی نفسیاتی برادران صوبےدار |
| اصطلاحات دستاویزی مولو یانہ لیفٹیننٹ متصوفانہ کوہستانی جمالیاتی معصومانہ |

**APPENDIX C. Urdu suffixes list**

|  |
| --- |
| ت ا ی ہ ھ س ء ؤ ح ر و ز ٹ ڑ ں ۃ ک |
| ۃً تو ات ونی نک یت وا رس گی یس سٹ اؤ یے وند نگ وں یز ہا سی یہ یا تا نہ شو رو جو بہ با بر بے چہ چی را اں ہں ڑا ڑی ری زن ئی  گر اک نا تر پن پا او ہٹ کش سا یو تی شا سہ تی تا دہ پن نے نا ئے ال ور ئی تا گو چی تے تی گر رہ اں یں ین کن رد چے وے ا ًوی لش ہی کن وں تہ ئش نی ان نا |
| اؤں تیں ئیں جات واس باش ئیو یین توں اری یچہ یئن فس نوش روی توز بین کشی آرا کشی زدا نگر راں کدے انے ویز وتی اتی خور گین نہا ندی لیا دیا آور یاس ئیت پنا گار وان الو واں ناک ساز پوش گیر حور چیں ریز سوز زدہ کشا بوس کار باز دوز شدہ بان حال جگر مند ئيہ طور ینی اپن اوٹ ئیے گان باز رنگ ئیس انگ کدہ داد نما یدہ وری کار لوش دار پنا تاؤ ساز کدے اہٹ یاب گیر زنی گری وار مندی بازی داں  کوش کدہ سرا زدی شکن بند خیز ئیہ گیو انہ ئیں سرا بخت یلا اسی ویں پتی گاہ یلے دان ینے اتے دست راز یوں تری طلب دلی ربا فام یار بخش |
| ئیوں نزاد منڈی یالی ٹائی آیند دستی آگیں خورہ بانی پسندی زدگی نویس گوئی گیری وائی کراں گراں آباد آمےز اتوں نامے ارتے وائز بندی خوری پوشی نامی دانی نشین ئیت سازی یناس کاری تانی آزما وانا خانہ ٹھاک تانی کاری گاری پنات پنیت پرست پذیر پسند یچوں شکنی ستان نامہ ئیاں جاتی یانا گوار دوست خواہ یابی فراز آزاد پناہ گانہ انتا اننا انتی گزار ٹاتی ائزڈ یتیں یتوں زادہ اروں ئگان ئگاں فہمی ہٹوں یاری یاتی ترین شناس آیند زدگی تامہ ورزی پرور آوری فروش بینی مقدس کدوں فشاں خوان گداز ناکی نشیں کردہ جوئی خیزی گزیں خواں داری رساں فکری خوار طراز گسار کوشی سوزی آلود گیوں نفسی دیدہ نگار زدوں چاری سنجی بختی رنگی وایا سویں ٹانہ گردی پراں ازمز کیسز نواز آمیز  آموز کناں نگیں ہندی گاہی خانے وانے |
| گواری بیانی آرائی کنندہ گزراں دوراں آزادی شکنوں نگاہی یافتہ انگیز ڈہال پتیاں افشاں ستانی اندیش دہندہ نمائی تھانی بیانی فشانی فروشی سرائی رسانی پروری نشینی گرافز اندوز پرداز بر یوں گر یوں سراؤں آلودہ آمیزی آرائی ماراں حیراں نگراں میراں باراں شبانہ ترانہ ہشانہ گرانہ خانوں گاروں دہانی دانوں گواری الدین مندان مثانہ افراز افروز پزیری روائی آفریں گاران انائی ہٹیں اعظم اعل ٰی کشائی زمانہ دارنی زدگان نویسی ماراں  روروں خواری شعاری معنوی طرازی بازوں آفرین نشینی مندوں خانوں گاہیں گاہوں خوانی گیروں داروں ثانیہ خارجہ شناسی گساری کشیدہ  رسیدہ انداز ورسٹی نگاری سرائی گاہیں گزاری شدگان پرستی |
| گار یاں پسندوں گواریت دار یوں خارجیہ دارانہ طفلانہ پروانے ورز یاں وائزرز گردانی برادری دلنشین گرافیز ٹیگیاں لیواؤں خیز یوں گاہئیں نویسوں گوئیوں سکنوں بوسیوں ساز یاں ریز یاں کار یوں رانیوں آفرینی بہاراں ءمراں اختراں نگاراں کارانہ بازانہ جو یانہ روزانہ مندانہ مستانہ غرضانہ افسانہ جرمانہ بیگانہ افزائی خداداد بردار لاہیاں ناترسی یفکیشن دستیوں دستیاں بینیاں خواروں پردازی اہٹیں انگیزی اندوزی باز یوں باز یاں کار یاں ناکیوں ناکیاں کثیرال بندیوں بندیاں خیز یوں خیز یاں پذیروں پذیری پارینہ خوانوں دہندوں دار یاں نویساں نگاروں دانیوں دانیاں دیوانہ گوئیوں گوئیاں  اندازی اسیویں سازانہ دوگانہ ترکانہ علمانہ الہانہ متحدال |
| پاشیدگی آہنگیوں وندکریم نخواستہ ناگہانی رسانیوں رسانیاں آرایانہ گزارانہ شامیانہ تاز یانہ سوفیانہ مجرمانہ پرور یاں انگریزی نواز یوں نواز یاں بیانیاں  فشانیاں اندازوں تراشیاں پرور یوں خوانیوں پسندانہ حاکمانہ مالکانہ پرستانہ شفاخانہ ظر یفانہ گسترانہ غائبانہ شاعرانہ آرایانہ پنجگانہ آگار یوں غاصبانہ خطیبانہ خودداری شعارانہ کشائیوں کشائیاں نویسیوں وندزادہ آرائیوں آرائیاں رسدخانہ |
| برادر یوں مخالفانہ العلمانہ سپردداری |

**APPENDIX D. Tokenization marker**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Symbol | Unicode | Description | Symbol | Unicode | Description |
| { | 007B | Left curly brackets | A | 0041 | English capital letter A |
| } | 007D | Right curly brackets | B | 0042 | English capital letter B |
| ( | 0028 | Left parenthesis | C | 0043 | English capital letter C |
| ) | 0029 | Right parenthesis | D | 0044 | English capital letter D |
| [ | 005B | Left square bracket | E | 0045 | English capital letter E |
| ] | 005D | Right square bracket | F | 0046 | English capital letter F |
| > | 003D | Greater than sign | G | 0047 | English capital letter G |
| < | 003C | Less than sign | H | 0048 | English capital letter H |
| - | 002E | Full stop | I | 0049 | English capital letter I |
| \_ | 005F | Low line | J | 004A | English capital letter J |
| = | 003D | Equal sign | K | 004B | English capital letter K |
| + | 002B | Plus sign | L | 004C | English capital letter L |
| ÷ | 00F7 | Division sign | M | 004D | English capital letter M |
| \* | 002A | Asterisk | N | 004E | English capital letter N |
| ! | 0021 | Exclamation mark | O | 004F | English capital letter O |
| | | 007C | Vertical bar | P | 0050 | English capital letter P |
| \ | 005C | Backslash | Q | 0051 | English capital letter Q |
| / | 002F | Slash (Solidus) | R | 0052 | English capital letter R |
| : | 003A | colon | S | 0053 | English capital letter S |
| ؛ | 06B | Arabic semi-colon | T | 0054 | English capital letter T |
| ؟ | 061F | Arabic question mark | U | 0055 | English capital letter U |
| ~ | 007E | Tilde | V | 0056 | English capital letter V |
| @ | 0040 | At the rate sign | W | 0057 | English capital letter W |
| # | 0023 | Number sign | X | 0058 | English capital letter X |
| $ | 00A4 | Currency sign | Y | 0059 | English capital letter Y |
| % | 0025 | Percentage sign | Z | 005A | English capital letter Z |
| & | 0026 | Ampersand sign | a | 0061 | English small letter a |
| ^ | 005E | Circumflex accent | b | 0062 | English small letter b |
| - | 002D | Hyphen -minus | c | 0063 | English small letter c |
| ، | 060C | Arabic comma sign | d | 0064 | English small letter d |
| ; | 003B | Semicolon | e | 0065 | English small letter e |
| “ | 201C | Left double quotation mark | f | 0066 | English small letter f |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ” | 201D | Right double quotation mark | g | 0067 | English small letter g |
| ‘ | 2018 | Left single quotation mark | h | 0068 | English small letter h |
| ’ | 2019 | Right Left single quotation mark | i | 0069 | English small letter i |
| ” | 0022 | Quotation mark | j | 006A | English small letter j |
| ! | 0021 | Exclamation mark | k | 006B | English small letter k |
| - | 06D4 | Arabic period (Dash) | l | 006C | English small letter l |
| ‘ | 0027 | Apostrophe | m | 006D | English small letter m |
| ۰ | 06F0 | Arabic digit zero | n | 006E | English small letter n |
| ۱ | 06F1 | Arabic digit one | o | 006F | English small letter o |
| ۲ | 06F2 | Arabic digit two | p | 0070 | English small letter p |
| ۳ | 06F3 | Arabic digit three | q | 0071 | English small letter q |
| ۴ | 06F4 | Arabic digit four | r | 0072 | English small letter r |
| ۵ | 06F5 | Arabic digit five | s | 0073 | English small letter s |
| ۶ | 06F6 | Arabic digit six | t | 0074 | English small letter t |
| ۷ | 06F7 | Arabic digit seven | u | 0075 | English small letter u |
| ۸ | 06F8 | Arabic digit eight | v | 0076 | English small letter v |
| ۹ | 06F9 | Arabic digit nine | w | 0077 | English small letter w |
| 0 | 0030 | European digit zero | x | 0078 | English small letter x |
| 1 | 0031 | European digit one | y | 0079 | English small letter y |
| 2 | 0032 | European digit two | z | 007A | English small letter z |
| 3 | 0033 | European digit three |  | 0x0D | Carriage return |
| 4 | 0034 | European digit four |  | 0x0A | Line feed |
| 5 | 0035 | European digit five |  | 0x0C | Form feed |
| 6 | 0036 | European digit six |  |  |  |
| 7 | 0037 | European digit seven |  |  |  |
| 8 | 0038 | European digit eight |  |  |  |
| 9 | 0039 | European digit nine |  |  |  |

