Find Similar Words (Using Paratext to Spell Check Indic scripts):

Note that many of the characters that need to be entered here, such as vowel signs and invisible characters, cannot be typed in isolation, and must be entered by typing the 4-digit codepoint and then pressing **Alt+x**.

Some of the “similar” characters below represent phonological differences in many languages (differences that can change the meaning of the word), while others don’t. It’s most productive to first focus on spelling differences that do not change the meaning (such as ignoring the zero-width joiner/non-joiner characters, or including स/श/ष if not phonologically distinct), and then afterwards add in pairs that are only similar.

Letters that Sound Alike:

Start with whichever pairs/sets that may get mixed, especially in borrowed terms/names:

स/श/ष न/ण ञ/न्य व/ब क/ख ग/घ च/छ ज/झ त/थ द/ध प/फ ब/भ त/ट थ/ठ द/ड ध/ढ अ/आ इ/ई ि/ी उ/ऊ ु/ू ए/ऐ े/ै ओ/औ ो/ो ं/ँ/ङ

(If you use the short vowels, such as ऐ and ओ, you may want to include those variants in these sets too.)

To include geminates (eg. अन्नि vs अनि), also include pairs like this:

क/क्क ख/क्ख ग/ग्ग घ/ग्घ ङ/ङ्ङ च/च्च छ/च्छ ज/ज्ज झ/ज्झ ट/ट्ट ठ/ट्ठ ड/ड्ड ढ/ड्ढ ण/ण्ण त/त्त थ/त्थ द/द्द ध/द्ध न/न्न प/प्प फ/प्फ ब/ब्ब भ/ब्भ म/म्म ल/ल्ल व/व्व स/स्स श/श्श ष/ष्ष

If ह or य or र ever alternate with something else, add pairs for those too. (e.g. If प्यार and पिअर are possible mixups, add ्या/िअ/ीअ and likewise for other vowels.)

Letters to ignore

In this setting, you definitely want to include the ZWJ and ZWNJ characters, as these are what make the difference between spellings like अन्न vs. अन्न vs. अन्न. These are invisible characters, but you can type them by keying their 4-digit hex code (200C and 200D) and then pressing **Alt+x**. Separate them with a space.

It may be worth including the nukta dot (़ 093C) here, if it ever occurs in your text.

It can be productive to include ा here (093E) so that it will group the inherent vowel /a/ with explicit /aa/. e.g. अन and अना.

When you consider variations like हात vs. हात्, because they are pronounced the same, it's tempting to add the final halant (094D) to the letters to ignore. Unfortunately, Paratext does not yet allow *positional rules* so that you can specify that only a *word-final* halant is optional. Ignoring all halants will cause your geminate rules to never match, and if you try to fix that by removing the halants from the geminate rules, you'll get all kinds of false

matches, such as grouping बानाना with बन. If you want to be thorough so as to catch the हात vs. हात् pairs, try adding the halant once you've been through without it, and look for anything new in what it finds.

Ignore Diacritics

This option is not intended for Indic users, so don't try to use it. (It's primary effect is to ignore any of the vowels that are marked above or below the consonant, but not any that are marked to the left or right of the consonant.)

Warning About Overlapping Sets

There are all kinds of phonological cases where alternation is likely to occur across one feature difference, but not across two. E.g. /b/ may alternate with /p/ (difference only of voicing), and /f/ may likewise alternate with /p/ (difference only of continuance), while /b/ does not alternate with /f/ (difference of both voicing and continuance).

You might reasonably expect that specifying separate sets of b/p and /p/f would ensure that the correct similarities are compared (and indeed, there are spell checking algorithms that would weigh single-feature differences as "better-matching" than multiple-feature differences), but unfortunately the way that Paratext currently implements this is to collapse this into b/p/f, with the result that false pairings of b/f will occur.

There is no simple workaround to this problem, but you might try running the tool first with one set and then afterwards with the other set.

Nasals

Nasals are a particularly tricky group even at the orthographic level, because there is often spelling alternation between using the candrabindu or anusvara marks (ं/ँ) and using the particular explicit nasal. For example, lots of languages have अंगूर and/or अँगूर and/or अङ्गूर, as they are all pronounced the same. It would be tempting to add ं/ँ/ङ् to the sound alike list, but ङ् is only one of the possible nasals. -The phonological value of candrabindu / anusvara is determined by the point of articulation of the following consonant.

The logical solution would be to specify a set for each nasal. (ं/ँ/ङ् ं/ँ/न् ं/ँ/म् etc.) Unfortunately, for the reasons explained above in the warning on overlapping sets, this will end up equating all nasals to each other.

The best workaround for now may be to generate a nasals set for every consonant, as in this sheet of Phonologically-Equivalent Nasals.

Devanagari:

ंक/ँक/ङ्क ंख/ँख/ङ्ख ंग/ँग/ङ्ग ंघ/ँघ/ङ्घ ंह/ँह/ङ्ह ंच/ँच/ञ्च ंछ/ँछ/ञ्छ ंज/ँज/ञ्ज ंझ/ँझ/ञ्झ ंय/ँय/ञ्य ंश/ँश/ञ्श ंट/ँट/ण्ट ंठ/ँठ/ण्ठ ंड/ँड/ण्ड ंढ/ँढ/ण्ढ ंर/ँर/ण्र ंष/ँष/ण्ष ंत/ँत/न्त ंथ/ँथ/न्थ ंद/ँद/न्द ंध/ँध/न्ध ंल/ँल/न्ल ंळ/ँळ/न्ळ ंस/ँस/न्स ंप/ँप/म्प ंफ/ँफ/म्फ ंब/ँब/म्ब ंभ/ँभ/म्भ ंव/ँव/म्व

Other Notes

The list of things to include will depend on the particular language. Some languages will have lots of legitimate minimal pairs with/without nukta or with/without aakar (कम vs काम in Nepali, for example), so whether to include those in the Ignore group will depend on the language. In the end it usually comes down to how many false matches you're willing to go through in order to be sure to catch every mistake.