Telugu Bigram Splitting using Consonant-based and Phrase-based Splitting

T. Kameswara Rao
Assoc. Professor and Head, CSE Dept
Brahma's Inst. of Engg. and Tech
Rajupalem, Nellore, AP, India

Dr. T. V. Prasad
Former Dean of Computing Sciences,
Visvodaya Technical Academy,
Kavali, AP, India

***Abstract*—Splitting is a conventional process in most of Indian languages according to their grammar rules. It is called '*pada vicchEdanam*' (a Sanskrit term for word splitting) and is widely used by most of the Indian languages. Splitting plays a key role in Machine Translation (MT) particularly when the source language (SL) is an Indian language. Though this splitting may not succeed completely in extracting the root words of which the compound is formed, but it shows considerable impact in Natural Language Processing (NLP) as an important phase. Though there are many types of splitting, this paper considers only consonant based and phrase based splitting.**

***Keywords*—Bigram; n-gram; consonant based splitting; phrase based splitting**

I. Introduction

Combining / conjunction of two or more words to form bigrams or n-grams is a conventional process in Indian languages which plays an important role [1], for instance, '*vibhakti'* (inflection) attachment to a root word that can be noun, pronoun, verb, etc. Inflections become postpositions and they are attached to the rear end of the root word. In many foreign languages like English, French, etc., inflections are prepositions and they are separate words. This is the reason why foreign languages strictly maintain word order. If the word order is changed, then the meaning of the sentence will be changed. For example, in the sentence 'Krishna is playing with snake', the word 'with' is preposition and related to 'snake'. The word order of 'with' and 'snake' should not be disturbed, and if changed, it may yield a sentence like 'snake is playing with Krishna'. Now, this sentence may be grammatically valid but gives incorrect meaning and objective of the sentence is changed.

But in most of the Indian languages, word order is negotiated [5] since there is no change in meaning as the inflections become part of the words. e.g. '*kRshNuDu pAmutO ADutunnADu*'. Here '*kRshNuDu*' is the Subject, '*pAmu*' is the Object, '*tO*' is the inflection, '*ADu*' is the verb and '*tunnADu*' is the tense and gender describer. Inflection and object are combined together to form single word using conjunction rules and meaning does not change whatever the word order may be. The above sentence can also be written with no change in meaning as:

1. *'pAmutO kRshNuDu ADutunnADu'*
2. *'pAmutO ADutunnADu kRshNuDu'*
3. *'ADutunnADu kRshNuDu pAmutO'*
4. *'ADutunnADu pAmutO kRshNuDu'*
5. *'kRshNuDu ADutunnADu pAmutO'*

This happens because of attaching inflection with root word i.e. '*tO*' with '*pAmu*'. If these two are not combined, word order affects the sentence considerably and may change the meaning or become meaningless. For example, '*kRshNuDu pAmu tO ADutunnADu*'. This can be written as '*pAmu kRshNuDu tO ADutunnADu*' (absurd meaning). Another example is '*rAmuDi valana rAvANuDi cAvu*' (*rAvaNa*'s death is due to *rAma*). Here '*valana*' is inflection and relates strictly with '*rAmuDi*'. Ignoring word order, in case this sentence is written as '*rAvANuDi valana rAmuDi cAvu*' (*rAma*'s death is due to *rAvaNa*) and the meaning is drastically changed.

All these examples conclude that when the inflections are properly attached to appropriate root words, then the word order cannot be an obligation, or else word order changes the meaning in a wrong direction and may render the sentence meaningless. The primary objective of the MT is to maintain the meaning. But in general, databases or dictionaries do not contain words with their inflection forms. As a consequence, splitting of those compounds is a mandatory step in MT to improve ease as well as accuracy in translation.

II. Issues in Conjunction and Splitting

Just as the issue of handling splitting leading to word order was discussed above, similar issue is faced with conjunctions also. Translators must be aware of when and where conjunctions are to be and not to be employed. If not, either unnecessary meanings are generated or sentence becomes either meaningless or non-informative. Two instances are given, one each for conjunction and splitting:

- ***Issues in conjunction:*** When the sentence. 'doMga rAmuNNi koTTADu' (literally meaning thief beat rAma) is examined, the word 'doMga' is subject and a noun, 'rAmuNNi' is inflected object and 'koTTADu' is verb and also acts as gender representative. There is nothing wrong in combining 'rAmuNNi' and 'koTTADu' to form a compound 'rAmuNNikoTTADu' from the meaning's perspective. Issues arise if the words 'doMga' and 'rAmuNNi' are combined to form the compound 'doMgarAmuNNi' which literally means 'thief natured rAma', a disturbed meaning. This happened since the noun 'doMga' is converted in to an adjective when it is combined with ‘rAmuNNi’. Moreover, since the sentence missed the subject, one cannot understand who has hit ‘rAma’.

- ***Issues in splitting:*** Examine the sentence ‘lakshmi piccikukkani caMpinadi’ (literally meaning lakshmi killed mad dog). Here ‘lakshmi’ is subject and is a noun, ‘kukkani’ is object and is inflected and ‘caMpinadi’ is verb as well as tense and gender representative. Splitting of ‘piccikukkani’ is to be done in such a way that it is to be considered as a whole word for translation. Otherwise, mere splitting in the sentence may result as ‘lakshmi picci kukkani caMpinadi’ (literally means lakshmi’s madness killed the dog) which is absurd though grammatically correct.

III. Sandhis As An Aid For Splitting

Amongst all Indian languages, Sanskrit and Telugu have well structured and numerous grammar rules [4]. Especially, the richness of Telugu language is with its huge number of words which can help express the meaning and mood much precisely [6]. Translation of compound-words (or simply compounds) or n-grams can be an obligation in MT as they are not available in database as they are. It becomes an overhead to maintain a database that consists of every possible conjunctional combination of n-grams. It is, therefore, impossible to translate compounds without splitting.

Though conjunction (known as ‘*sandhi*’ in Sanskrit, as well as in Telugu) is seemed to be combining of two words (*pUrva-pada and uttara-pada*), actual ‘*sandhi*’ occurs between only two letters, i.e. last letter of the first word (‘*pUrva-svara’*) and first letter of the second word (‘*para-svara*’). A *sandhi* will results in at least one of the following

- Concatenation of pUrva-pada and para-pada
- Either pUrva-svara or para-svara is dropped
- A new vowel / consonant is inserted
- Some specific words ***are inserted***

This paper handles only consonant and specific word issues.

‘*sandhis*’ are categorized in to five types in Sanskrit. They are 1. ‘*ach sandhis*’ 2. ‘*prakRti bhAva sandhis*’ 3. ‘*hal sandhis*’ 4. ‘*visarga sandhis*’ and 5. ‘*svAdi sandhis*’.

Among all these categories, only ‘*hal sandhis*’ are considered in this paper as they involves necessarily a consonant (consonant is called ‘*hal*’ in Sanskrit) in compound as a result. These ‘*hal sandhis*’ are listed in table 1.

TABLE I. Sanskrit ‘*Hal Sandhis*’ and Their Resultant Consonants

S.No	‘*sandhi*’ name	Resultant Consonant
1	*Scutva sandhi*	*S, c, ch, j, Q*
2	*shTutva sandhi*	*sh, T*
3	*jastva sandhi*	*g, j, D, d, b*
4	*anunAsika sandhi*	*G, Q, N, n, m*
5	*pUrva savarNa sandhi*	*ggh, jjh, DDh, ddh, bbh*
6	*para savarNa sandhi*	*Ll*
7	*chatva sandhi*	*Cch*

Though Telugu adopted all ‘*sandhis*’ from Sanskrit grammar, it has its own ‘*sandhis*’ as well as their precise formulae. Some of them involve only vowels [7], some of them involve only consonants and some involves both vowels and consonants. Later two cases are briefly considered as consonant resultant ‘*sandhis’* in this paper. Table 2 describes the list of Telugu consonant ‘*sandhis*’.

There are some other ‘*sandhis’* which works with phrases. They are discussed in later sections.

TABLE II. Telugu ‘*Hal Sandhis*’ and their resultant consonants

S.No	‘*sandhi*’ name	Resultant Consonant
1	*yaDAgama sandhi*	*y*
2	*dviruktaTakAra sandhi*	*TT*
3	*gasaDadavAdESa sandhi*	*g, s, D, d, v*
4	*druta / saraLAdESa sandhi*	*g, j, D, d, b*
5	*pumpvAdESa sandhi*	*p / Mp*
6	*penvAdi sandhi**	*nn*
7	*AmrEDita sandhi*	*TT / rr / ll/ tt*
8	*pampavarNAdESa sandhi*	*Pa*
9	*trika sandhi*	Two consonants other than *S,sh,s,h*
10	*lu la na la sandhi*	Two consonants
11	*dugAgama sandhi*	*du*
12	*nakArAdESa sandhi*	*nn / NN*

*This ‘*sandhi*’ is listed in phrase based splitting also

IV. Consonant Based Splitting

Consonant based splitting is not a special kind of splitting rather finding the possibilities to split a compound with the help of a consonant. When the consonants which are listed in Table 2 are encountered in a compound, splitting the compound by using appropriate ‘*sandhi’* rules yields good results in extracting root words with which the compound is formed. Except in some special cases, in majority of cases, this process is successfully extracted the root words of the compound. This paper deals with splitting rather than ‘*sandhi*’ formation. In view of this, deep explanation about ‘*sandhi’* or compound formation is negotiated.

1) yaDAgama sandhi: This ‘sandhi’ involves the consonant ‘y’ as a result in compound. (See Table 3)

TABLE III. Examples of *yaDAgama sandhi*

S.No	Root words	Compound	Replace ‘y’ with
1	*mA + amma*	*mA**y**amma*	A whitespace
2	*mI + illu*	*mI**y**illu*	A whitespace
3	*mA + Uru*	*mA**y**Uru*	A whitespace
4	*hari + ataDu*	*hari**y**ataDu*	NA
5	*tella + Enugu*	*tella**y**Enugu*	NA

When a ‘***y***’ is observed in the compound, Ensuring the previous vowel of the consonant ‘*y*’ is a long vowel (in this case ‘*A*’ or ‘*I*’) then applying *yaDAgama sandhi* rules in splitting yields good results in extracting root words.

Splitting: In this case root words can be extracted by replacing ‘*y*’ with a whitespace (i.e. removal of ‘*y*’). Splitting has to be done before ‘*y*’ and ‘*y*’ is replaced with a whitespace where second root word starts.

2) dviruktaTakAra sandhi: This ‘sandhi’ involves the consonant ‘TT’ as a result in compound. dviruktaTakAra *literally means the letter 'Ta' is twice derived. Examples are given in Table 4.*

TABLE IV. EXAMPLES OF *DVIRUKTATAKARA SANDHI*

S.No	Root words	Compound	Replace 'TT' with
1	*kuru + usuru*	*ku**TT**usuru*	*ru* + whitespace
2	*ciru + aDavi*	*ci**TT**aDavi*	*ru* + whitespace
3	*kaDu + eduru*	*ka**TT**eduru*	*Du* + whitespace
4	*naDu + aDavi*	*na**TT**aDavi*	*Du* + whitespace
5	*niDu + Urpu*	*ni**TT**Urpu*	*Du* + whitespace
6	*naDu + illu*	*na**TT**illu*	*Du* + whitespace

When '*TT*' is observed in the compound, *dviruktaTakAra sandhi* rules are applied in splitting to extract root words.

Splitting: '*TT'* can be suitably replaced with '*ru / Du*' and split. There are two special cases in this '*TT*' issue, as mentioned in Table 5 and 6.

TABLE V. SPECIAL CASE 1 OF HANDLING '*TT*'

S.No	Root words	Compound	Replace with
1	*ciTTi + eluka*	*ci**TT**eluka*	*TTi* + whitespace
2	*ciTTi + aDavi*	*ci**TT**aDavi*	*TTi* + whitespace
3	*ciTTi + Amudamu*	*ci**TT**Amudamu*	*TTi* + whitespace
4	*ciTTi + Idu*	*ci**TT**Idu*	*TTi* + whitespace
5	*ciTTi + uDuku*	*ci**TT**uDuku*	*TTi* + whitespace

*In the above mentioned special cases, '*TT*' can be replaced with '*TTi*' and split.

TABLE VI. SPECIAL CASE 2 OF HANDLING '*TT*'

S.No	Root words	Compound	Replace 'TT' with
1	*ciTTi + cApa*	*ci**TT**icApa*	NA
2	*ciTTi + pApa*	*ci**TT**ipApa*	NA
3	*ciTTi + tALamu*	*ci**TT**itALamu*	NA

In the above case, '*parasvara*' is consonant. In a compound, '*parasvara*' is neither available nor identifiable. For this case, vowel based splitting gives better results. Replacement for '*TT*' is not applicable in this case, rather separating '*ciTTi*' as a root word. Splitting has to be done after 'ru/Du/TTi' where first root word ends.

Issue 1: '*TT*' based splitting fails in the case of '*ceTTekkaDa'* (literally means where tree is). It is known that its root words are '*ceTTu' + 'ekkaDa'*. But according to Table 4, it become '*ceDu' + 'ekkaDa'* (literally means where bad is) which gives an incorrect meaning after translation.

This can be avoided by applying vowel based splitting [7] which gives longest word among all combinations of root words, i.e '*ceTTu*' and '*ceDu*' can be formed according to split rules. But '*ceTTu*' is longer word than '*ceDu*' and is returned as a first root word.

Issue 2: This '*sandhi*' rules fail in splitting of the compound '*miTTamadhyAhnamu*', since it is formed with an unusual logic by combining the root words '*madhyAnamu' + 'madhyAnamu*'.

N.B: Better idea to translate this category of words is to consider them as a single word rather bigram and enter them in to database, e.g. *'ciTTaDavi, ciTTeluka, naTTaDavi.*

Note: More issues are discussed in Phrase based splitting.

1) gasaDadavAdESa sandhi: This 'sandhi' involves the consonant 'g/s/D/d/v' as a result in compound. (Table 7)

TABLE VII. EXAMPLES OF *GASADADAVADESA SANDHI*

S.No	Root words	Compound	Replace with
1	*rAru + kadA*	*rAru**g**adA*	A whitespace + *k*
2	*apuDu + caniye*	*apuDu**s**aniye*	A whitespace + *c*
3	*nIvu + Takkari*	*nIvu**D**akkari*	A whitespace + *T*
4	*mIru + talaci*	*mIru**d**alaci*	A whitespace + *t*
5	*vAru + pOduru*	*vAru**v**Oduru*	A whitespace + *p*
6	*ataDu + kalaDu*	*ataDu**g**alaDu*	A whitespace + *k*
7	*vADu + cEsenu*	*vADu**s**Esenu*	A whitespace + *c*
8	*Ame + tolagenu*	*Ame**d**olagenu*	A whitespace + *t*

When '*g/s/D/d/v*' is identified in the compound, *gasaDadavAdESa sandhi* rules are applied in splitting to extract root words.

Splitting: In this case, root words can be extracted by simply replacing '*g/s/D/d/v*' with '*k/c/T/t/p*' appropriately with a whitespace prefixed. Splitting has to be done before '*g/s/D/d/v*' where second root word starts.

Issue: '*tallidamDrulu*' can be split according to this '*sandhi'* rule by replacing '*d*' with a whitespace + '*t'*. But in the case of '*Akalidappulu*' (literally means hungry and thirsty) this sandhi fails in giving correct meaning after translation as it splits the word into '*Akali' + 'tappulu*'(literally means hungry and mistakes). This also fails when it try to split the compound '*cerukugaDa*'(literally means stem of sugarcane). It splits it into '*ceruku + kaDa*'(literally means near sugarcane / at the end of the sugar cane) which is an incorrect translation.

Same problem is repeated in the case of '*nOrujAru*' which is formed by two root words '*nOru, jAru*' (logically means tongue slip). If this rule is applied on this it splits as '*nOru, cAru*' (literally means juice of mouth, i.e. spit) which is an incorrect translation.

2) druta / saraLAdESa sandhi: This 'sandhi' involves the consonant 'g/j/D/d/b' as a result in compound. (Table 8)

TABLE VIII. EXAMPLES OF DRUTA / SARALADESA SANDHI

S.No	Root words	Compound	Replace with
1	*pUcenu + kamala*	*pUcenu**g**amala*	whitespace + *k*
2	*Kanenu + cukkalu*	*Kanenu**j**ukkalu*	whitespace + *c*
3	*cEsenu + Takkulu*	*cEsenu**D**akkulu*	whitespace + *T*
4	*namilenu + tamba*	*Namilenu**d**amba*	whitespace + *t*
5	*virisenu + padmam*	*Virisenu**b**admam*	whitespace + *p*
6	*nannu + cUci*	*nannu**j**Uci*	whitespace + *c*
7	*bhAryanu + cEse*	*bhAryanu**j**Ese*	whitespace + *c*
8	*kappanu + tine*	*Kappanu**d**ine*	whitespace + *t*

When 'g/j/D/d/b' is identified in the compound, *druta / saraLAdESa sandhi* rules are applied in splitting to extract root words.

Splitting: In this, root words can be extracted by simply replacing 'g/j/D/d/b' with 'k/c/T/t/p' appropriately with a whitespace prefixed. Splitting has to be done before 'g/j/D/d/b' and the consonant is replaced (as in Table 8) where second root word starts.

Issues: This sandhi can form the compounds in different ways. They are listed in Table 9.

TABLE IX. TYPES OF *DRUTA/ SARALADESA SANDHI* COMPOUND FORMATIONS

S.No	Root words	Compound	Replace with
1	*pUcenu + kamala*	*pUceM**g**amala/ pUcen**g**amala*	WS+ *k*
2	*kanenu + cukkalu*	*kaneM**j**ukkalu /kanen**j**ukkalu*	WS+ *c*
3	*cEsenu + Takku*	*cEseM**D**akkulu/cEsen**D**akku*	WS+ *T*
4	*namilenu + tamba*	*namileM**d**amba/namilen**d**amba*	WS+ *t*
5	*virisenu + padma*	*viriseM**b**adma /virisen**b**adma*	WS+ *p*
6	*nannun + cUci*	*nannuM**j**Uci /nannun**j**Uci*	WS+ *c*
7	*bhAryan + cEse*	*bhAryan**j**Ese /bhAryaM**j**Ese*	WS+ *c*
8	*kappan + tine*	*kappan**d**ine /kappaM**d**ine*	WS+ *t*

*WS stands for whitespace

N.B: To improve accuracy in splitting, it is to be ensured that if the previous vowel of '*g/j/D/d/b*' is '*M/n*' abefore applying *druta/ saraLAdESa sandhi* rule. If so, then first replace '*M/n*' with '*nu*' and then replace the consonant according to Table 9.

Issue 1: This '*sandhi'* rules fails in splitting the compound '*vEsenugAlamu*'. It is known that its root words are '*vEsenu + gAlamu*'(literally meaning anchored). But according to '*sandhi'* rules, root words become *'vEsenu + kAlamu'* (literally meaning time is thrown), an incorrect translation.

Issue 2: If the compound is '*vaccenugOvulu*' formed by the words '*vaccenu, gOvulu*' (literally meaning cows came). '*sandih*' rules changes '*gOvulu*' into '*kOvulu*' and then searches in Database which is not a proper word and is not available. As a consequence, the compound cannot be split. One more example for this type of failure is '*pADenugandharvuDu*' formed by '*pADenu, gandharvuDu*' (literally meaning '*gandharva'* sung).

Note: Sometimes '*pUrvapada*' is a past tense of a verb which cannot be a root word and is not available in Database.

E.g. '*pUcenu'* (blossomed), '*kanenu'* (saw / delivered), '*cEsenu'* (done). Their root words are '*pUyu*' (To blossom), '*kanu*' (To see / To deliver), '*cEyu*'(To do) respectively. Sometimes '*pUrva pada*' is terminated with a '*druta*'(half M i.e. *n,* e.g. *pucen*)[2] which is not available in Database.

If a morphological algorithm is developed to morph the word '*pUcenu*' into '*pUyu*' and '*nannun*' into '*nannu*' and so on properly, then this '*sandhi*' rules are suitable for splitting.

1) puMpvAdESa sandhi: This 'sandhi' involves 'pu/Mpu' as a result in compound. Examples are given in Table 10.

TABLE X. EXAMPLES OF *PUMPVADESA SANDHI*

S. No	Root words	Compound	Replace with
1	*sarasamu + mATa*	*sarasa**pu**mATa/sarasa**Mpu**mATa*	*mu* + W
2	*virasamu + cUpu*	*virasa**pu**cUpu /virasa**Mpu**cUpu*	*mu* + W
3	*nikkamu + nIlamu*	*nikka**pu**nIlamu /nikka**Mpu**nIlamu*	*mu* + W

*W stands for whitespace

When a '*pu/Mpu'* is observed in the compound, *puMpvAdESa sandhi* rules are applied for splitting.

Splitting: In this case root words can be extracted by replacing '*pu/Mpu*' with '*mu*' with a whitespace appended. Splitting has to be done next to '*mu'* after '*pu/Mpu*' replaced with '*mu*' appended with a whitespace where second root word starts.

2) penvAdi sandhi: This 'sandhi' involves 'nn' as a result in compound. (Table 11)

TABLE XI. EXAMPLES OF *PENVADI SANDHI*

S.No	Root words	Compound	Replace with
1	*penu + adurulu*	*Pe**nn**adurulu*	*nu* + whitespace
2	*kanu + Aku*	*ka**nn**Aku*	*nu* + whitespace
3	*anu + urvISuDu*	*a**nn**urvISuDu*	*nu* + whitespace
4	*penu + oDalu*	*pe**nn**oDalu*	*nu* + whitespace

If '*nn*' is observed in the compound, *penvAdi sandhi* rules are applied in splitting to extract root words.

Splitting: In this case root words can be extracted by replacing '*nn*' with 'n*u*' with a whitespace appended. Splitting has to be done next to '*nu'* after '*nn*' replaced with '*nu*' appended with a whitespace where second root word starts.

3) AmrEDita sandhi: This 'sandhi' also involves the consonant 'TT' as a result in compound like dviruktaTakAra sandhi. But this 'sandhi' forms compounds in three types but this paper considers only one type as the second type does not involve consonant as a result and third type is more ambiguous and the nature is unidentifiable which is discussed as a special issue in this concept. (Table 12)

TABLE XII. EXAMPLES OF *AMREDITA SANDHI*

S.No	Root words	Compound	Replace with
1	*kaDa + kaDa*	*Ka**TT**akaDa*	Handled with a different logic.
2	*civara + civara*	*Ci**TT**acivara*	
3	*eduru + eduru*	*E**TT**aeduru*	
4	*naDuma + naDuma*	*na**TT**anaDuma*	
5	*pagalu + pagalu*	*pa**TT**apagalu*	
6	*bayalu + bayalu*	*ba**TT**abayalu*	
7	*modaTa + modaTa*	*mo**TT**amodaTa*	
8	*tuda + tuda*	*Tu**TT**atuda*	
9	*kona + kona*	*Ko**TT**akona*	
10	*piDugu + piDugu*	*pi**TT**apiDugu*	

Case 1: When '*TT*' is observed in the compound then, search for the immediate next consonant in the compound not allowing to surpass two vowels. If only one vowel is placed between '*TT*' and its next immediate consonant, then check immediate previous consonant in the compound. If both are same then, extract a word starting from the next immediate consonant to the end of the compound; and identify the compound is formed of two words of that extracted word kind. See all examples of Table 12 except 3.

Case 2: If two vowels are passed in finding the next immediate consonant in compound, then identify the second vowel and identify the immediate previous vowel in the compound. If both are same, then extract a word starting from second vowel to the end of the compound; and identify the compound is formed of two words of that extracted word kind. See the third example in Table 12.

Issues: Special issues are given in Table 13.

TABLE XIII. SPECIAL ISSUES OF *AMREDITASANDHI*

S.No	Root words	Compound	Replace with
1	*iMkulu + iMkulu*	*irriMkulu*	NA
2	*iggulu + iggulu*	*irriggulu*	NA
3	*ceduru + ceduru*	*cellaceduru*	NA
4	*tuniyalu + tuniyalu*	*Tuttuniyalu*	NA
5	*miTlu + miTlu*	*mirumiTlu*	NA
6	*atuku + atuku*	*aMdatuku*	NA
7	*iMkulu + iMkulu*	*irriMkulu*	NA
8	*tumuru + tumuru*	*Tuttumuru*	NA

*NA stands for Not-Applicable

If these compounds are tried to split using this sandhi rules, they cannot extract proper root words as well as proper translation. However, if some rules are implemented to split above compounds, then there are very high chances of improper splitting of other compounds which are not formed based on this sandhi rule. For example, a rule is developed to split '*cellaceduru*' into '*ceduru + ceduru*' then it splits the compound '*tellateppalu*' which is formed of '*tella, teppalu*' (literally means white boats) into '*teppalu + teppalu*' which is an incorrect splitting.

NB: Entry of above category of words as it is in to the Database gives better results than splitting.

1) paMpavarNAdESa sandhi: This 'sandhi' involves 'pa' as a result in compound. Examples are given in Table 14.

TABLE XIV. EXAMPLES OF *PAMPAVARNADESA SANDHI*

S.No	Root words	Compound	Replace with
1	*nAmu + cEnu*	*nA**pa**cEnu*	*mu* + whitespace
2	*pAmu + rEDu*	*pA**pa**rEDu*	*mu* + whitespace
3	*janumu + nAra*	*janu**pa**nAra*	*mu* + whitespace
4	*vEmu + kAya*	*vE**pa**kAya*	*mu* + whitespace
5	*ammu + Sayya*	*am**pa**Sayya*	*mu* + whitespace
6	*inumu + kaDDi*	*inu**pa**kdaDDi*	*mu* + whitespace
7	*enumu + guMpu*	*enu**pa**guMpu*	*mu* + whitespace
8	*minumu + gAre*	*minu**pa**gAre*	*mu* + whitespace
9	*emmu + gUDu*	*eM**pa**gUDu*	*mu* + whitespace
10	*kanumu + cEnu*	*kanu**pa**cEnu*	*mu* + whitespace

When a '*pa*' is observed in the compound, *paMpavarNAdESa sandhi* rules are applied .

Splitting: In this case, root words can be extracted by replacing '*pa*' with 'm*u*' with a whitespace appended. Splitting has to be done next to '*pa'* after '*pa*' replaced with '*mu*' appended with a whitespace where second root word starts.

Issues: This '*sandhi'* rules fail when applied to split the compound '*pApakannu*' which is formed by two root words '*pApa, kannu*' (literally means baby's eye). '*pAmu, kannu*' (literally means snake's eye) will be the root words after splitting the compound using this rules which is an incorrect translation.

2) trika sandhi: This 'sandhi' involves two same consonants other than 'S,sh,s,h' as a result in compound. In Sanskrit grammar, 'a, i, e' are called 'trika'[3]. (See Table 15)

TABLE XV. EXAMPLES OF *TRIKA SANDHI*

S.No	Root words	Compound	Replace with
1	*A + kanya*	*A**kk**anya*	A precise procedure is used rather than mere replacement
2	*I + kAlamu*	*i**kk**Alamu*	
3	*E + lOkamu*	*e**ll**Okamu*	
4	*A + aSvamu*	*a**yy**aSvamu*	
5	*A + bhaMgi*	*a**bbh**aMgi*	
6	*I + adi*	*I**yy**adi*	
7	*I + dharaNi*	*i**ddh**araNi*	

When any two consonants other than '*S,sh,s,h*' are observed in the compound as in the above table, *trika sandhi* rules are applied in splitting to extract root words.

N.B: To apply this rule, ensure the immediate preceding character of two consonants is any one of the vowels '*a/i/e*'. Otherwise this rule is not applicable.

Splitting: In this case, root words are to be extracted by replacing first consonant along with its preceding short vowel with the long form of the vowel. Splitting has to be done between two consonants.

3) lu-la-na-la sandhi: This 'sandhi' involves two same consonants as a result in compound. This 'sandhi' is also called 'Du-varNalOpa sandhi' as it eliminates the rear consonant 'Du' of 'purvapada' in compound. In this case 'purvapada' is mandatorily 'mUDu'. (Table 16)

TABLE XVI. EXAMPLES OF *LU-LA-NA-LA SANDHI*

S.No	Root words	Compound	Replace with
1	*mUDu + jagamulu*	*mu**jj**agamulu*	First consonant of the two is replaced with '*Du'*
2	*mUDu + lOkamulu*	*mu**ll**Okamulu*	
3	*mUDu + Arulu*	*mu**yy**Arulu*	
4	*mUDu + maDugu*	*mu**mm**aDugu*	
5	*mUDu + kAru*	*mu**kk**Aru*	
6	*mUDu + pAtika*	*mu**pp**Atika*	

If two consonants are observed in the compound as in the above table, *lu-la-na-la sandhi* rules are applied if and only if the two preceding letters of two consonants are '*mu*'.

Splitting: In this case, root words are to be extracted by replacing first consonant with '*Du*'. Splitting has to be done between two consonants.

4) dugAgama sandhi: This 'sandhi' involves 'du' as a result in compound. Examples are given in Table 15.

TABLE XVII. EXAMPLES OF *DUGAGAMA SANDHI*

S.No	Root words	Compound	Replace with
1	*nI + karuNa*	*n**Idu**karuNa*	'*du*' is replaced with a white space
2	*nA + nEramu*	*nA**du**nEramu*	
3	*tana + rUpu*	*tana**du**rUpu*	
4	*mana + sAyamu*	*mana**du**sAyamu*	
5	*tama + karuNa*	*tama**du**karuNa*	
6	*nA + rUpu*	*nA**du**rUpu*	
7	*mI + cUpu*	*m**Idu**cUpu*	
8	*mA + snEhamu*	*mA**du**snEhamu*	

When 'du' is observed in the compound, *dugAgama sandhi* rules are applied in splitting to extract root words.

Splitting: In this case, root words are to be extracted by replacing '*du*' with a whitespace.

5) nakArAdESa sandhi: This 'sandhi' involves two same consonants either 'nn' or 'NN' as a result in compound. Examples are given in Table 18.

TABLE XVIII. EXAMPLES OF NAKARADESA SANDHI

S.No	Root words	Compound	Replace with
1	*mUDu + nALLu*	*mU**nn**ALLu / mU**NN**ALLu*	Different logic is applied
2	*reMDu + nALLu*	*re**nn**ALLu / re**NN**ALLu*	

When '*nn/NN*' observed in the compound as in the above table, *nakArAdESa sandhi* rule is applied in splitting to extract root words.

Splitting: In this case, root words are to be extracted by replacing first '*n/N*' of the pattern '*nn/NN*' with '*MDu'* + a whitespace. If the second consonant of the pattern is '*N*', then, change it in to '*n*'. Split two consonants.

Issues: This '*sandhi*' rule fails in yielding proper words in the case of '*vENNILLu*'(literally meaning hot water), a compound formed by two proper words '*vEDi, nILLu*'. This rule can split the compound as '*vEDu, nILLu*' (literally meaning requesting water) which is an improper splitting.

V. PHRASE BASED SPLITTING

Here the phrase is referred to be the '*pUrvapada'.* The '*sandhis*', like '*rugAgama sandhi'* of Telugu, directly deal with phrases instead of single vowel or consonant. If the phrase is identified, compound can be split by applying these '*sandhi*' rules. But over dependency on these rules leads to improper results. These '*sandhis'* are listed in Table 19.

TABLE XIX. LIST OF PHRASE BASED '*SANDHIS'*

S. No	'*sandhi*'	*sandhi* Phrases
1	*rugAgama sandhi*	*pEda, bIda, mudda, bAliMta, manuma, goDDu, komma, vidhava, dhIra and some words which ends with 'aMta' like SrImaMta, guNavaMta, dhImaMta, bhAgyavaMta,* etc.
2	*prAtAdi sandhi*	*prAta, krotta, lEta, pUta, pUvu, mIdu, muMdu, keMpu, cennu, kriMdu,* etc.
3	*penvAdi sandhi*	*penu, kanu, anu,* etc
4	*lu – la – na – la sandhi*	*mUDu*
5	*dviruktaTakAra sandhi*	*kuru, ciru, kaDu, niDu, naDu*

*4 and 5 '*sandhis'* are listed in '*hal sandhis'* section also.

In this phrase based '*sandhis*', phrase is '*pUrva-pada*' in majority of cases. In many instances, these '*sandhis'* does not result in the '*pUrva-pada*' as it is in the compound. For this reason, it is required to create a new phrase as an independent phrase or as a reference of '*pUrva-pada*'. A search is done for this phrase in compound for better results. Table 20 a,b,c, d & e describes the possibility of phrase based splitting.

TABLE XX. (A). POSSIBILITIES OF '*RUGAGAMA SANDHI'*

S.No	Root words	Compound	Phrase searched for	A/NA
1	*pEda + Alu*	*pEdarAlu*	*arAlu*	A
2	*bIda + Alu*	*bIdarAlu*	*arAlu*	A
3	*bAliMta + Alu*	*bAliMtarAlu*	*arAlu*	A
4	*goDDu + Alu*	*goDDurAlu*	*urAlu*	NA
5	*komma + Alu*	*kommarAlu*	*urAlu*	NA

Splitting: According to grammar rules, the '*para-pada*' of this sandhi is always '*Alu*'. For better results, ensuring the presence of the new phrase '*arAl/urAl*' in the compound is essential. Then, separate the root word from the compound as in the Table 20 (a). Next, as per rules, the second part's first letter i.e. '*r*' is to be removed to make it '*Alu*'. As the word '*Alu*' have two meanings – 'wife and lady', replacing '*r*' with a meaningless character (it becomes *@Alu*/ *#Alu / *Alu etc.*) avoids the ambiguity in the selection of the suitable meaning of '*Alu*'. For this, a rule is to be implemented to pick the word 'lady' from the database when '*Alu*' is preceded with a special symbol. If not, it becomes 'poor wife' instead of 'poor lady'.

In the case of '*dhanavaMturAlu/ guNavaMturAlu/ dhIrurAlu / bhAgyavaMturAlu*' etc. replace '*u*' with '*a*' of '*urAl*', and separate the root word. Remaining process can be same. These cases give improper meaning after splitting. Because '*goDDurAlu*' literally means a married lady who have no children. It can be merely split into '*goDDu*' + '*Alu',* when this rule is applied. This is acceptable according to grammar rule, but the meaning becomes 'buffalo lady'. Though ordinary splitting is applied on this compound, it can be split in to '*goDDu*' + *'rAlu*' literally means 'buffalo to fell' which is a misinterpretation. Better way to solve these splitting problems is, to include these compounds in database not considering them as bigrams.

TABLE XX. (B). POSSIBILITIES OF '*PRATADI SANDHI'*

S.No	Root words	Compound	Phrase searched	A/NA
1	*kriMdu + kaDupu*	*krIgaDupu*	*krI*	A
2	*kriMdu + kannu*	*krIgannu*	*krI*	A
3	*kriMdu + kAlu*	*krIgAlu*	*krI*	A
4	*kriMdu + toDa*	*krIdoDa*	*krI*	A
5	*krotta + cAya*	*kroMjAya*	*kroM*	A
6	*krotta + pasiDi*	*kroMbasiDi*	*kroM*	A
7	*krotta + mAvi*	*krommAvi*	*kro**	NA
8	*prAta + illu*	*prAyillu*	*prA**	NA
9	*lEta + dUDa*	*lEdUDa*	*lE**	NA
10	*pUvu + tOTa*	*pUdOTa*	*pU**	NA
11	*mIdu + kaDa*	*mIgADa*	*mI*	NA
12	*krotta + gaMDi*	*kroggaMDi*	*kro**	NA

Splitting: If the first three characters of a compound are '*krI',* then, they are to be replaced with '*krimda*' and if first four characters of a compound are '*kroM*', then they are to be replaced with '*krotta*' if and only if the immediate next character of the phrase are '*g/j/D/d/b*'. For these two cases. change it to '*k/c/T/t/p*' respectively.

Sometimes '*g/j/D/d/b*' is also included into the phrase to improve the quality in splitting. For example, '*muMgOpamu*' is the bigram formed by the root words '*muMdu' + 'kOpamu*'. For this case, phrase will be '*muMg*'. But only '*muM*' is replaced with '*muMdu*' and '*g*' *(/j/D/d/b*) of the phrase is changed to '*k*' (*/c/T/t/p*). Remaining process is same.

**Issues:* Though they are acceptable according to the grammar rules, sometimes they mislead the proper words. For example, '*prAraMbhasamayamu*' is the compound (literally means starting time). It can become '*prAta' + 'raMbhasamayamu'* (literally means the time of old *raMbha* – a divine dancer) when split using this rule where the phrase used is '*prA*' as in the above table. This is incorrect translation.

TABLE XX. (C). POSSIBILITIES OF '*PENVADI SANDHI*'

S.No	Root words	Compound	Phrase searched for	A/NA
1	*penu + aduru*	*Pennaduru*	*Penn*	A
2	*penu + Uta*	*pennUta*	*Penn*	A
3	*kannu + Aku*	*kannAku*	*kann*	A
4	*penu + Eru*	*pennEru*	*penn*	A

Splitting: If the first four characters of a compound are '*penn/kann*' then replacing with '*penu/kannu*' is the phrase based splitting using '*penvAdi sandhi'* as in the above table.

Issues: '*pennidhi*' (literally means great treasure) is a compound formed by two root words '*penu'*+ *'nidhi*'. For this case '*n*' is prefixed to the second word after splitting.

TABLE XX. (D). POSSIBILITIES OF '*LU- LA- NA- LA SANDHI*

S.No	Root words	Compound	Phrase to be searched for	A/NA
1	*mUDu + jagamulu*	*mujjagamulu*	*mujj*	A
2	*mUDu + yEDulu*	*muyyEDulu*	*muyy*	A
3	*mUDu + lOkamulu*	*mullOkamulu*	*mull*	A
4	*mUDu + pAtika*	*muppAtika*	*mupp**	A
5	*mUDu + kArulu*	*mukkArulu*	*mukk**	A
6	*mUDu + ciccu*	*mucciccu*	*mucc**	A
7	*mUDu + trIva*	*muttrOva*	*mutt*	A
8	*mUDu + cemaTalu*	*muccemaTalu*	*mucc**	A

Splitting: When first two characters are '*mu*' and next two characters are any same consonants, then '*mu*' is replaced with '*mUDu*' and separated. First character from the second word i.e. first consonant is removed as in the above table.

**Issues:* Sometimes a root word like '*muccaTa*' (literally means fondness) can also be split into '*mUDu' + 'caTa*' (an absurd) lead to incorrect translation.

TABLE XX. (E). POSSIBILITIES OF '*DVIRUKTATAKARA SANDHI'*

S.No	Root words	Compound	Phrase searched for	A/NA
1	*kuru + usuru*	*ku**TT**usuru*	*kuTT**	A
2	*ciru + aDavi*	*ci**TT**aDavi*	*ciTT**	A
3	*kaDu + eduru*	*ka**TT**eduru*	*kaTT**	A
4	*naDu + aDavi*	*na**TT**aDavi*	*naTT**	A
5	*niDu + Urpu*	*ni**TT**Urpu*	*niTT**	A

Some descriptions are given about this in previous section.

**Issues:* This sort of compounds highly conflicts with '*AmrEDita sandhi*' as in the Table 21.

TABLE XXI. COMPOUNDS OF DVIRUKTATAKARA AND AMREDITA SANDHIS

S.No	Compounds of *dviruktaTakAra sandhi* rule	Compounds of *AmrEDita sandhi* rule
1	*ka**TT**eduTa(kaDu + eduTa)*	*ka**TT**akaDa(kaDa + kaDa)*
2	*ci**TT**aDavi(ciru + aDavi)*	*ci**TT**acivara(civara + civara)*
3	*na**TT**aDavi (naDu + aDavi)*	*na**TT**anaDi(naDi+ naDi)*
4	*ni**TT**Urpu (niDu + Urpu)*	*ni**TT**aniluvu(niluvu + niluvu)*

Some issues of '*dviruktaTakAra sandhi'* are described above.

As the '*pUrva-pada*' does not appear in the compound except some portion, the compounds formed by the *AmrEDita sandhi* rule also looking same and is not possible to decide which rule is to be used to split. To avoid unnecessary issues raised in splitting the compounds, enter them in database as they are.

VI. CONCLUSION

Though there are many issues involved in consonant and phrase based splitting that lead to incorrect translation, these rules extracts root words from the bigrams considerably. Many problems can be solved either by implementing a rule which finds out the longest root word among many combinations or entry of critical compounds in to database. For instance, the compound '*muppYokaTi*' (literally means thirty one) formed by root words '*muppY'* + *'okaTi*'. It can be split as '*mUDu*'+ *'pYokaTi*'. If further splitting process continues, the word '*muppY*' will be available in database as it is a root word, and it can be separated. This becomes '*muppY'* + *'okaTi*'. As the word '*muppY'* is longer than '*mupp'*, it becomes the result.

REFERENCES

[1] Malladi Krishna Prasad, "Telugu Vyaakaranamu", Sri Venkateswara Book Depot, 2012.

[2] Dr. Samudrala Vemkata Ramga Ramanujacharya, "Samskruta Vaani" Rohini Publications, 1997.

[3] Kambhampati Ramagopala Krishnamurti, "Telugu Vyaakaranamu", Sri Sailaja Publications, 1991.

[4] A.H. Arden, "A Progressive Grammar of the Telugu Language", 2nd Edition, Society for promoting Christian Knowledge, Madras, 1905.

[5] Robert Caldwell, "A Comparative Grammar of The Dravidian or South-Indian Family of Languages", 2nd Edition, 1875.

[6] T. Venkateswara Prasad and G. Mayil Muthukumaran Telugu to English Translation Using Direct Machine Translation Approach, Int. J. of Sc. & Engg. Investigations, Vol.2 Issue 12, Jan 2012

[7] T. Kameswara Rao and Dr. T.V. Prasad, " Key Issues in Vowel based Splitting of Telugu Bigrams", Int. J. of Adv. Computer Sc. App., Vol. 5, No. 3, 2014

APPENDIX

Pronunciation: Letter pronunciation is as in Table 20 and 21.

TABLE XXII. VOWEL PRONUNCIATION

RT	Usage as	RT	Usage as	RT	Usage as
a	a in – That	R	Ru in – Ruk	o	O in - Obey
A	a in - Father	Ru	roo in – roof	O	oa in - Roar
i	i in - His	e	e in - When	W	ou in - Shout
I	Ea in - Eagle	U	oo in – fool	aM	um in - sum
u	u in – Put	E	a in - Hate	aH	aH in – aH
U	oo in – fool	Y	I in - Ice		

RT means Roman Telugu

TABLE XXIII. CONSONANT PRONUNCIATION

RT	Usage as	RT	Usage as	RT	Usage as
k	C in – Cut	Dh	Dh in – Dhamaru	y	Y in – Yak
kh	che in – Ache	N	Na in – Jana gaNa mana	r	R in – Rat
g	g in - Dog	t	th in – Path	l	L in – Lip
gh	gh in -Vagha	th	th in - sthambh	L	l in – mahila
G	Gy in - Gyan	d	Th in – The	v	V in – Van
c	ch in – Catch	dh	dh in – dharma	S	S in – Sand
Ch	Ch in – Chunk	n	n in – Pen	sh	Sh in - Sharp
J	J in – Jar	p	P in – Pot	h	H in – Hen
Jh	Jh in - Jhaveri	ph	Ph in – Phal		
T	t in – Cat	b	B in – Bat		
Th	T in - Tagore	bh	bh in – prabhu		
D	d in – God	M	M in - Mat		

The capitalized letters should be pronounced with greater emphasis on them.