Key Issues in Vowel Based Splitting of Telugu Bigrams

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 of compound Telugu words into its components or root words is one of the important, tedious and yet inaccurate tasks of Natural Language Processing (NLP). Except in few special cases, at least one vowel is necessarily involved in Telugu conjunctions. In the result, vowels are often repeated as they are or are converted into other vowels or consonants. This paper describes issues involved in vowel based splitting of a Telugu bigram into proper root words using Telugu grammar conjunction ('sandhi') rules for MT.**

Keywords—***Telugu word splitting; vowel based splitting; compound word splitting; bigrams; trigrams; n-grams; NLP***

I. Introduction

Sanskrit is considered as the mother language for almost all Indian languages, since a majority of the Indian languages are based on grammar rules similar to that of Sanskrit grammar [6]. Sanskrit is grammatically very well structured and very rich in its inflections [7]. It is the oldest language on the earth to have a powerful structured grammar. Panini (300 BCE) the greatest grammarian developed Sanskrit grammar with more than 4000 rules [8], [10]. Unlike western languages, Sanskrit is the best example that unites the words to form a compound word (or simply compound). According to Bloomfield and Chomsky (1957), sentence is the largest grammatical unit [16].

There is a possibility and custom to write a complete sentence as a single compound in Sanskrit. For instance "*jalObhaumamantarikshamitidvidhAbhavati*" – for convenience, it can be tokenized as "*jalaH bhaumaM antarikshaM iti dvidhA bhavati*" means 'water is of two types, one is on the earth, and another is in space' ('*jalaH*' – water, '*bhaumaM*' – on the earth, '*antarikshaM*' – in space, '*iti*' – like this, '*dvidhA*' – two categories, '*bhavati*' – is).

Sanskrit scholars are to be very careful about tokenization. Lack of appropriate knowledge on the grammar or less attention to each and every letter gives immature tokenization that leads to yield distorted or quite opposite meaning in some special cases [8]. For example '*viSvAmitraH*' is the word to be tokenized; its meaning is friend of the universe. It can be tokenized as '*viSva'* + *'amitraH*' according to '*savarNadIrga sandhi*', which is not to be applied here because it gives opposite meaning i.e., enemy of the universe. For this kind of special cases, Sanskrit gives exemptions strictly. So it should be '*viSva'* + *'mitraH*', where regular conjunction rule is to be violated and special rule is applied. The person who is aware of this kind of special cases can only tokenize properly.

Likewise majority of Indian languages follow the features of Sanskrit; undergo conjunction which is inevitable that lead in generating compounds that are essentially bigrams, trigrams or n-grams. Bigram is a compound formed by two words and trigrams by three words, and so on. As Telugu is one of them, one can see the nature of uniting the words to form n-grams in Telugu also. Though Telugu is highly influenced by other languages, especially most of it is by Sanskrit [7], Telugu is not originated from Sanskrit [4]. Even though Telugu was originally intended to be totally free from Sanskrit, it has tremendous impact and deep penetration into Telugu. In 1816, Francis White Ellis raised this issue. Later Bishop Robert Cardwell proved that a family of twelve Dravidian languages Telugu, Tamil, Kannada/Canarese, Malayalam, Tulu, Kodagu/Coorg, Tuda, Kota, Gond, Khond/Ku, Rajmahal and Oraon are not originated from Sanskrit in his book titled "A Comparative Grammar of Dravidian Languages" in 1856 [5]. As a proof of that, pure Telugu literature work is available in the form of *'yayAti caritramu*' by '*ponnagaMTi telaganna*' written in 16th century [1][13]. Later Telugu mingled with Sanskrit heavily by '*samskrutAndhra kavulu*' (Sanskrit – Andhra - a synonym of Telugu - poets) when they translated epics in Sanskrit literature like Ramayana, Mahabharata and *bhAgavata*, etc. Learning or speaking Sanskrit was a great honor in those days and literature work in Sanskrit was highly honored. That can be one of the reasons to Sanskritize Telugu to enhance its value.

Additionally, there are numerous dialects in Indian languages - even many of them do not have script and are based on their culture, territory, and have tremendous impact of non-Indian languages like Urdu, Persian, Arabic, English, etc. For instance, most of the Telugu language is affected by Urdu in '*telaMgANa*' territory. '*tarfIdu, aafIsu, pennu, pEparu, kaburlu, bassu'*, etc. are words from those languages adapted in Telugu [4]. Such words, their conjunctions and their corrupted / colloquial forms are almost understandable by local humans but not easily by non-locals. For example '*nI jimmaDa*' is the word very frequently used by the natives of eastern Andhra. It means 'let your tongue fall' (literally 'jimma' is the colloquial form of '*jihva*' – Sanskrit word for tongue, '*aDa*' is the corrupted form of '*paDu*' – a Telugu word).

II. VOWELS

According to ‘*pANini*’ Sanskrit grammar, vowels and their forms are given as in TABLE I [2] (pronunciations are given in Appendix).

TABLE I. CHARACTERISTICS OF VOWEL ‘*A* (अ)’

Vowel		Time required to pronounce	Types
Roman English	***dEva nAgari***		
a (short vowel)	**अ** (*hrasva*)	One unit (*Eka mAtra*)	*anudAtta, udAtta, svarita*
A (Long vowel)	**आ** (*dIrgha*)	Two units (*dvi mAtra*)	*anudAtta, udAtta, svarita*
*A*3 (Longer vowel)	**आ3** (*pluta*)	Three units (*tri mAtra*)	*anudAtta, udAtta, svarita*

Note: ‘*pluta*’ is applied in calling somebody who is at a distance. For example, ‘*hE rAmA3*’. Here ‘3’ indicates the ‘*pluta*’ of the vowel ‘*A*’. If ‘*pluta*’ is not applied here, the person cannot be called.

Again each type is classified in to two different forms, namely ‘*anunAsika*’ and ‘*ananunAsika*’. ‘*anunAsika*’ is a nasal sound while ‘*ananunAsika*’ is not. A total of six types for each vowel ‘*a, A, A*3’ yields 18 different forms of vowel ‘*a*(अ)’. Likewise ‘*i*(इ), *R*(ऋ)’ also have 18 different forms each. ‘*z*(ऌ), *E*(ए), *Y*(ऐ), *O*(ओ), and *W* (औ)’ can be obtained in 12 forms for each as they are not derived long forms. A huge total of 132 vowels are there in Sanskrit. Mostly these are used in ‘Vedas’. But only thirteen vowels ‘*a, A, i, I, u, U, R, Ru, z, E, Y, O* and *W*’ are used in general usage. Two more vowels ‘*aM (anusvAra), aH (visarga)*’ are used in Sanskrit. Two special vowels are there appears only in Sanskrit named ‘*jihvamUlIyam’* and ‘*upadhmAnIyam’*. If *‘visarga’* is appeared as prior character of consonant ‘*k*’, it is considered as ‘*artha-visarga’* and called ‘*jihvamUlIyam’*, e.g. ‘*aMtaHkaraNam’*. If ‘*visarga*’ is appeared as prior character of consonant ‘*p*’, it is called ‘*upadhmAnIyam*’. Ex. ‘*vAyuHpaMkam*’.

Telugu includes two more short vowels ‘*e*’ and ’*o*’ and one more long vowel ‘*Z*’ to the above listed Sanskrit vowels to comprise a total of eighteen vowels [2]. All proper Telugu words end with vowels only. That’s why Telugu language is called ‘*ajanta*’ (= ‘*ach*’ + ‘*anta*’, literally ‘*ach*’ meaning vowel and ‘*anta*’ meaning ending) language. Consonants are called ‘*hal*’ in Telugu. They are 37 in number. Unlike Telugu, words of almost all Indian languages end in consonants and hence called ‘*halanta*’ languages. All western languages are also categorized as ‘*halanta*’ languages as their words commonly end in consonants except Italian that ends in vowels. This is the reason why Telugu is called ‘Italian of the East’ and one of the secrets behind sweetness of Telugu vocabulary.

There are eighteen vowels in Telugu language as shown in TABLE II. All the vowels are called ‘*ach*’ or ‘*svara*’ according to Telugu grammar, their Roman equivalents are as in TABLE II.

TABLE II. TELUGU VOWELS AND THEIR ROMAN EQUIVALENT

Telugu vowel	**అ**	**ఆ**	**ఇ**	**ఈ**	**ఉ**	**ఊ**	**ఋ**	**ౠ**	**ఌ***
Roman English	*A*	*A*	*I*	*I*	*U*	*U*	*R*	*Ru*	*Z*
Telugu vowel	**ౡ***	**ఎ**	**ఏ**	**ఐ**	**ఒ**	**ఓ**	**ఔ**	**అం**	**అః**
Roman English	*Z*	*E*	*E*	*Y*	*O*	*O*	*W*	*aM*	*aH*

Note: *Vowels ‘**ఌ, ౡ** (*z, Z*)’ are not used now-a-days, they are not considered in this paper.

In Telugu, vowels are classified into two types as follows [14].

- ‘*hrasvAs’* – *‘a, i, u, R, z, e’*
- *‘dIrghAs’* – *‘A, I, U, Ru, Z, E, Y, O, W’*

‘*dIrghAs*’ again classified into two types as follows.

- ‘*vakrAs’* – *‘e, E, o, O’*
- *‘vakratamAs’* – *‘Y, W’*

III. PROCESS OF SPLITTING WORDS

Due to many practical issues involved in maintaining a database with all combinations of compounds, it is better to maintain only standard or root words. Compounds of the source language are split to obtain the original words using reverse engineering in accordance to the conjunction (‘*sandhi*’) rules. This will make the morphological analysis easier.

Proper stemming and correcting of corrupted forms for splitting of n-grams into individual tokens is necessary for better understanding the context. This plays an important role in translation also whereas understanding is also a kind of translation. Splitting of compounds into root words is an important phase in NLP for the applications like MT [9]. Building a computational model to analysis natural language is the goal of NLP [15]. For MT from Telugu to any other language including Indian languages, one of the issues of dealing with source language words is that each word need to be stored in the database together with different suffixes/prefixes (also known as inflections) thus tremendously increasing the storage space. This is in the case like Telugu that has about 800 different dialects within the state of Andhra Pradesh. But most of the conjunctions are common and are computable. The best way to translate them is to split back into root words as they formed and then translate individual root words. Compounds are formed with two or more root words. While the root words can be retrieved from database, the inflections thus obtained needs serious focus. Inability to sufficiently handle the inflections may result in false word formations and distorted meaning. But mere splitting the compound may not give complete meaning all the time. To understand the meaning of a compound, first identify the meaning of components and then the relationship between them [11]. For instance, a compound ‘*rAmunitOkapirAju*’ is formed by two words ‘*rAmunitO + kapirAju*’.

First word is inflected, and second word is a root word. '*rAmunitO*' literally means with '*rAma*' and '*kapirAju*' means Hanuman. If the inflection is not observed in the first word, it may be split as '*rAmunitOka*' (literally means the tail of '*rAma*') + '*pirAju*' (an absurd word), which gives a distorted meaning.

The scope of this paper is limited to deal with bigrams only for obtaining better MT and aims to propose solutions to the issues of vowel based splitting. Issues related to handling of different types of dialects and their corrupted forms have not been considered. More specifically, handling of compounds formed according to the grammar rules, and their splitting based on vowels together with certain special cases in Telugu have been discussed.

IV. Conjunction Rules

Splitting is a process opposite to the conjunction. Conjunction is called '*sandhi*' and splitting is called '*sandhi vicchEda*' in Sanskrit. Telugu also use the word '*sandhi*' to represent conjunction. At least two words are required for conjunction. First word is called '*pUrva pada*' and the second word is called '*para pada / uttara pada*' [12]. While most part of the word remains unchanged, technically, the actual '*sandhi*' occurs between two letters, i.e., '*pUrva svaram*'(last letter of the '*pUrva pada*') and '*para svaram*' (first letter of the '*para pada*') [1]. Telugu language adapted many of the '*sandhi*' rules from Sanskrit as it uses much grammar of Sanskrit in addition to its own grammar rules. Sanskrit grammar rules were adapted into Telugu since majority words of Telugu language were taken from Sanskrit. Sanskrit grammar describes three ways to form a '*sandhi*'. They are

- ***'Agamamu'*** (literally means coming in Sanskrit): one new letter comes according to '*sandhi*' rules, and is included between the conjunction characters, without removing any of them. Ex. '*mA' +' amma' = 'mAyamma'*. '*A'* and '*a*' are involved in '*sandhi*', the new letter '*y*' is included between '*A*' and '*a*'. '*tut*' is introduced in '*tuDAgama*', '*dud*' in '*dhuDAgama*', '*jam*' in '*jamuDAgama*' and so on, are the examples of '*Agama sandhis*' in Sanskrit and '*yaDAgama, TugAgama, rugAgama*' etc. are the examples of '*Agama sandhis*' in Telugu.
- ***'AdESamu'*** (literally means rule in Sanskrit): one new letter replaces the two '*sandhi*' letters. Ex. '*parama' + 'ISvaruDu' = 'paramESvaruDu'.* '*a*' and '*I*' are involved in '*sandhi'* and both are replaced with '*E*'. '*yaNAdESa, anunAsika*' etc., are the examples of '*AdESa sandhis*' in Sanskrit and '*pumpvAdESa, gasaDadavAdESa*' etc., are the examples of '*AdESa sandhis*' in Telugu.
- ***'EkAdESamu'*:** one character of the '*sandhi*' letters are omitted and second one continues to exist in the compound. Ex. '*rAmuDu' + 'ataDu' = 'rAmuDataDu'*. '*u*' and '*a*' are involved in '*sandhi*', but letter '*u*' is dropped and only '*a*' is continued. '*savarNadIrgha, guNa, vRddhi*' etc., are the examples of '*EkAdESa sandhis*' in Sanskrit, and '*akAra, ukAra, ikAra sandhis*' are the examples in Telugu.

In Sanskrit, there are five important classifications of '*sandhis*'. They are 1) '*ach sandhi*', 2) '*hal sandhi*' 3) '*visarga sandhi*' 4) '*prakruti bhAva sandhi*' and 5) '*svAdi sandhi*' [1]. But only first three '*sandhis*' are used very frequently. '*ach*' and '*visarga sandhis*' works with vowels and '*hal sandhis*' works with consonants. '*sandhi*' classifications are given in TABLE III.

TABLE III. List Of Sanskrit '*sandhis*'

***sandhi* Type**	**Names**
'ach'	*savarNa dIrgha, guN, vRddhi, yaNAdESa, vAntAdESa, yAntAdESa, pUrva rUpa, para rUpa, avaJnAdESA*
'hal'	*Scutva, shTutva, jaStva, anunAsika, pUrva savarNa, para savarNa, chatva*
'*visarga*'	This '*sandhi*' shows six types of differences, but names are not given to them.

Though these three kinds of '*sandhis*' are used by Telugu as it is, they are treated as Sanskrit '*sandhis*'. Telugu defines around thirty '*sandhis*' (TABLE IV) according to its grammar. These Telugu '*sandhis*' fall under '*ach sandhis*', '*hal sandhis*' or work with both vowels as well as consonants [1].

TABLE IV. List Of Telugu '*sandhis*'

S.No	***'sandhi' name***	***S.No***	***'sandhi' name***
1	*ukAra (utva)*	*16*	*penvAdi*
2	*yaDAgama*	*17*	*AmrEDita*
3	*AkAra*	*18*	*muvarNalOpa*
4	*IkAra*	*19*	*paDvAdi*
5	*apadAdisvara*	*20*	*aligAgama*
6	*dvirukta TakAra*	*21*	*anukaraNa*
7	*TugAgama*	*22*	*visandhi*
8	*RugAgama*	*23*	*paMpavarNAdESa*
9	*gasaDadavAdESa*	*24*	*trika*
10	*saraLAdESa(druta)*	*25*	*lu-la-na-la*
11	*puMpvAdESa*	*26*	*dugAgama*
12	*pugAgama*	*27*	*allOpa*
13	*prAtAdi*	*28*	*nakArAdESA*
14	*nugAgama*	*29*	*mivarNalOpa*
15	*itvAdESa*	*30*	*ukAra vikalpa sandhi*

Though there are many Sanskrit and Telugu '*sandhis*', only some of them for vowel based splitting have been considered which are resulting in a vowel in compound (TABLE V) irrespective of they are classified as '*ach sandhi', 'hal sandhi*' or '*visarga sandhi*'. Some special cases are also discussed in this paper even they are involving a consonant.

TABLE V. List Of '*sandhis*' Results A Vowel In Compound

S.No	**'*sandhi*' name**	**Result vowel**	***S/T***
1	*savarNadIrgha*	*A,I,U,Ru*	S
2	*guNa*	*E,O,ar*	S
3	*vRddhi*	*Y, W*	S
4	*visarga*	*O, H*	S
5	*akAra, ikAra, ukAra*	*a,A,i,I,u,U,e,E,Y,o,O,W*	T
6	*yaNAdESa*	*y* + vowel	T
7	*jastva sandhi*	*g/j/D/d/b* + vowel	S
8	*dviruktaTakAra*	*TT* + vowel	T

*S – Sanskrit, T – Telugu

1) 'savarNa dIrgha sandhi': This results a vowel 'A/I/U/Ru' accordingly when one of the following (TABLE 6) pattern occurs.

Note: Pattern is the combination of '*purvasvara*' and '*parasvara*'

TABLE VI. ALL PATTERNS OF '*SAVARNADIRGHA SANDHI*'

S.No	Pattern	Res	Example
1	a + a	A	phAla + aksha = phAl**A**ksha
2	a + A	A	rAma + Alayamu = rAm**A**layamu
3	A + a	A	pUjA + arhuDu = pUj**A**rhuDu
4	A + A	A	prajA + Anati = praj**A**nati
5	i + i	I	kavi + iMdruDu = kav**I**MdruDu
6	i + I	I	naMdi + ISvara = naMd**I**Svara
7	I + i	I	vANI + iMdra = vAN**I**Mdra
8	I + I	I	vasumatI + ISa = vasumat**I**Sa
9	u + u	U	su + ukti = s**U**kti
10	u + U	U	mRdhu + Uruvu = mRdh**U**ruvu
11	U + u	U	vadhU + unnati = vadh**U**nnati
12	U + U	U	vadhU + Uruvu = vadh**U**ruvu
13	R + R	Ru	pitR + RNamu = pit**Ru**Namu
14	R + Ru	Ru	Examples are not given since no words start or end with 'Ru' in Telugu.
15	Ru + R	Ru	
16	Ru + Ru	Ru	

2) 'guNa sandhi': This results in a vowel 'E/O/ar' accordingly when one of the following (TABLE VII) pattern occurs.

TABLE VII. ALL PATTERNS OF '*GUNA SANDHI*'

S.No	Pattern	Res	Example
1	a + i	E	bhUtala + itara = bhUtal**E**tara
2	a + I	E	svarga + ISuDu = svarg**E**SuDu
3	A + i	E	mahA + ikshu = mah**E**kshu
4	A + I	E	mahA + ISuDu = mah**E**SuDu
5	a + u	O	dAma + udara = dAm**O**dara
6	a + U	O	Nava + Uha = nav**O**ha
7	A + u	O	mahA + uttama = mah**O**ttama
8	A + U	O	mahA + UrU= mah**O**rU
9	a + R	ar	brahma + Rshi = brahm**ar**shi
10	A + R	ar	mahA + Rshi = mah**ar**shi

3) 'vRddhi sandhi': This results a vowel 'Y/W' accordingly when one of the following (TABLE VIII) pattern occurs.

TABLE VIII. ALL PATTERNS OF '*VRDDHI SANDHI*'

S.No	Pattern	Res	Example
1	a + E	Y	Eka + Eka = Ek**Y**ka
2	a + Y	Y	Sarva + YSvarya = sarv**Y**Svarya
3	A + E	Y	kAMtA + Eka = kAMt**Y**ka
4	A + Y	Y	mahA + YSvarya = mah**Y**Svarya
5	a + O	W	Eka + Oshadhi = Ek**W**shadhi
6	a + W	W	rAma + Wnnatya = rAm**W**nnatya
7	A + O	W	mahA + Odhana = mah**W**dhana
8	A + W	W	kAMtA + Wnnati = kAMt**W**nnati

4) 'visarga sandhi': This has five rules of which only two are considered since these two rules results in a vowel 'O/H' accordingly when one of the following (TABLE IX) pattern occurs.

Rule1: when '*pUrva svara*' is '*aH*' and '*para svara*' is '*a/u/g/gh/j /jh/D/Dh/d/dh/n/b/bh/m/y/r/l/v/h*', then '*pUrva svara*' is replaced with '*O*' in the compound.

TABLE IX. ALL PATTERNS OF '*VISARGA SANDHI*- 1'

S.No	Pattern	Res	Example
1	aH + a	O	saH + ahaM = s**O**haM
2	aH + u	O	vijayaH + ullAsa = vijay**O**llAsa
3	aH + g	O	tiraH + gamana = tir**O**gamana
4	aH + gh	O	manaH + ghana = man**O**ghana
5	aH + j	O	saraH + ja = sar**O**ja
6	aH + jh	O	manaH + jhari = man**O**jhari
7	aH + D	O	naraH + DiMbha = nar**O**DiMbha
8	aH + Dh	O	SivaH + Dhamar = Siv**O**Dhamar
9	aH + d	O	yaH + dEvaH = y**O**dEvaH
10	aH + dh	O	tapaH + dhana = tap**O**dhana
11	aH + n	O	yaSaH + nagara = yaS**O**nagara
12	aH + b	O	manH + buddhi = man**O**buddhi
13	aH + bh	O	manaH + duHkh = man**O**duHkh
14	aH + m	O	SiraH + maNi = Sir**O**maNI
15	aH + y	O	manaH + yaMtra = man**O**yaMtra
16	aH + r	O	rajH + rAgamu = raj**O**rAgamu
17	aH + l	O	jalaH + lahari = jal**O**lahari
18	aH + v	O	tapaH + vanaM = tap**O**vanaM
19	aH + h	O	manaH + hara = man**O**hara

Note: In this case, '*visarga*' should be preceded by '*a*' else, this rule is not applicable. Ex. '*dhanuH*' + '*chalanamu*' = '*dhan**uSc**alanamu*'.

Rule2: *H + k / kh / p / ph* gives '*visarga*' as it is in the compound (TABLE X).

TABLE X. ALL PATTERNS OF '*VISARGA SANDHI*-2'

S.No	Pattern	Res	Example
1	H + k	Hk	tapaH + kaMpa = tapa**Hk**ampa
2	H + kh	Hkh	hariH + khaDga= hari**Hkh**aDga
3	H + p	Hp	dhanuH + puMja= dhanu**Hp**uMja
4	H + ph	Hph	SaSiH + phalamu = SaSi**Hph**alamu

Note: For this rule any vowel can precede the '*visarga*' and that vowel appears in the compound with preceding character of '*visarga*'.

5) 'ukAra sandhi': If 'pUrva svara' is 'u' and 'paras vara' is a vowel, then 'u' is replaced by the vowel in result (TABLE XI).

TABLE XI. ALL PATTERNS OF '*UKARA SANDHI*'

S.No	Pattern	Res	Example
1	u + a	a	iTlu + anenu = iTl**a**nenu
2	u + A	A	kAlu + ADu = kAl**A**Du
3	u + i	i	vADu + ippuDu = vAD**i**ppuDu
4	u + I	I	kAlu + IDcu = kAl**I**Dcu
5	u + u	U	nEDu + unnADu = nED**u**nnADu
6	u + U	U	mEmu + Ugamu = mEm**U**gamu
7	u + e	E	Enugu + ekku = Enug**e**kku
8	u + E	E	vAgu + EtAmu = vAg**E**tAmu
9	u + Y	Y	siddhamu + Y = siddham**Y**
10	u + o	o	rAmuDu + okaDu = rAmuD**o**kaDu
11	u + O	O	ippuDu + Orpu = ippuD**O**rpu
12	u + W	W	tinu + WshadhaM= tin**W**shadhaM
13	u + M	M	ipuDu + aMtamu = ipuDa**M**tamu

Note: if '*ukAra sandhi*' rule is applied to split '*vAgISuDu*', it becomes '*vAgu' + 'ISuDu*', which is a wrong splitting. It should actually be split as '*vAk' + 'ISuDu*'. Such conflicts should be handled carefully and may require manual checks.

6) 'akAra sandhi': This 'sandhi' has four rules but only one of them is considered since remaining results in a consonant.

Rule: when '*pUrva svara*' is 'a' and '*parasvara*' is any vowel, then '*a*' is replaced by the vowel in result (TABLE XII).

TABLE XII. ALL PATTERNS OF '*AKARA SANDHI*'

S.No	Pattern	Res	Example
1	*a + a*	*a*	*rAma + anna = rAm**a**nna*
2	*a + A*	*A*	*ciMta + Aku = ciMt**A**ku*
3	*a + i*	*I*	*puTTina + illu = puTTin**i**llu*
4	*a + I*	*I*	*cinna + Iga = cinn**I**ga*
5	*a + u*	*u*	*cUDaka + uMDu = cUDaku**M**Du*
6	*a + U*	*U*	*Kotta + Uyala = kott**U**yala*
7	*a + e*	*e*	*sIta + ekkaDa = sIt**e**kkaDa*
8	*a + E*	*E*	*tella + Enugu = tell**E**nugu*
9	*a + Y*	*Y*	*nava + YSvarya = nav**Y**Svarya*
10	*a + o*	*o*	*cIma + okaTi = cIm**o**kaTi*
11	*a + O*	*O*	*konta + Opika = kont**O**pika*
12	*a + W*	*W*	*maha + WnnatyaM= mah**W**nnatyaM*

7) 'ikAra sandhi': if 'pUrva svara' is 'i' and 'parasvara' is a vowel, then 'i' is replaced by the vowel in result (TABLE XIII).

TABLE XIII. ALL PATTERNS OF '*IKARA SANDHI*'

S.No	Pattern	Res	Example
1	*i + a*	*a*	*Emi + aMTivi = Em**a**MTivi*
2	*i + A*	*A*	*nallani + Avu = nallan**A**vu*
3	*i + i*	*I*	*vacciri + ipuDu = vaccir**i**puDu*
4	*i + I*	*I*	*ciTTi + ItakAya = ciTT**I**takAya*
5	*i + u*	*u*	*idi + unnadi = id**u**nnadi*
6	*i + U*	*U*	*cakkani + Uru = cakkan**U**ru*
7	*i + e*	*e*	*idi + evaridi = id**e**varidi*
8	*i + E*	*E*	*Takkari + Enugu = Takkar**E**nugu*
9	*i + Y*	*Y*	*idi + YrAvatamu = id**Y**rAvatamu*
10	*i + o*	*O*	*nETiki + okkaTi = nETik**o**kkaTi*
11	*i + O*	*O*	*idi + Orugallu = id**O**rugallu*
12	*i + W*	*W*	*ciTTi + Wshadhi = ciTT**W**shadhi*

Note: There are some special issues in this 'sandhi', like 'cEsi' + 'ipuDu' = 'cEsi**yi**puDu', 'vacci' + 'iccenu' = 'vacci**yi**ccenu'.

V. VOWEL BASED SPLITTING RULES

Technically, whatever the rules used for conjunction, they are used in reverse order to obtain those root words back. This approach can be considered as a reverse engineering process.

Algorithm:

1) A compound in Telugu, which is to be translated, is taken and is transliterated into Roman Telugu.

2) Each character is checked to determine whether it is a vowel.

3) If it is a vowel, then try all possible combinations to split the word according to the 'sandhi' rules listed in Tables 6 through 13.

4) If the compound is formed according to 'sandhi' rules of two words, then it is split into two words.

5) The process is recursively processed till all the words thus separated are found in the dictionary/database.

Example: '*SivArcana*' – formed by the root words '*Siva' + 'arcana*'. While using vowel based splitting, the vowels of '*SivArcana*' i.e., '*i, A, a*' are to be checked (TABLE XIV).

TABLE XIV. POSSIBLE PATTERNS OF THIS SPLITTING OF '*SIVARCANA*'

V	Pattern	Sandhi	Result	NA/A
i	*u + i*	*ukAra*	*Su + ivArcana*	NA
i	*a + i*	*akAra*	*Sa + ivArcana*	NA
i	*i + i*	*ikAra*	*Si + ivArcana*	NA
A	*a + a*	*savarNadIrgha*	*Siva + arcane*	**A**
A	*a + A*	*savarNadIrgha*	*Siva + Arcana*	NA
A	*A + a*	*savarNadIrgha*	*SivA + arcane*	NA
A	*A + A*	*savarNadIrgha*	*SivA+ Arcana*	NA
A	*u + A*	*ukAra*	*Sivu + Arcana*	NA
A	*a + A*	*akAra*	*Siva + Arcana*	NA
A	*i + A*	*ikAra*	*Sivi + Arcana*	NA
a	*u + a*	*ukAra*	*SivArcu + ana*	NA
a	*a + a*	*akAra*	*SivArca + ana*	NA
a	*i + a*	*ikAra*	*SivArci + ana*	NA

Note: If V (vowel) is the last character of the compound, then there will be a chance of split when the letter is a long vowel like '*A, I, U, E, Y, O' e.g., 'vaccADA' = 'vaccADu' + 'A*' (means, 'did he come?').

This is occurs almost in interrogative cases. But there is no chance for short vowels to be the result of conjunction. There is no need to check the last character of the compound, if it is '*a, i, u, R, e or o*' assuming it is the result of '*sandhi*'. From all the patterns listed in TABLE XIX, '*a + a*' pattern of '*savarNadIrgha sandhi*' is applicable to split '*SivArcana*' into '*Siva + arcana*'. Amongst these 13 patterns, only one pattern is suitable to split the compound properly.

When one pattern splits the compound successfully, then there is no need to go for further splitting until unless the compound is formed by three or more. Unnecessary splitting may yield improper or unacceptable root words. As a rule of thumb, best results are obtained by splitting in such a way that first word extracted from the compound is as long as possible. Even if a proper word is obtained from the compound much before finishing, splitting process is not to be stopped until all vowels of the compound are checked. Ex. '*adhikAramaDugu*' is a bigram formed by two proper words '*adhikAramu*' (authority) and '*aDugu*' (to ask) by the rule of '*ukAra sandhi*'. But it can also be assumed as a trigram formed by three proper words '*adhi*' (to overcome), '*kAramu*' (chilli powder), '*aDugu*' (to ask). If splitting process is stopped at the earlier stage when it found a proper word (for instance, '*adhi*'), it yields useless or distorted meaning when translated.

Sometimes, some words are not to be treated as compounds and should be translated as a whole. For instance, '*adhikAri*' literally meaning "officer" is the word to be treated as single word and should not be split. If it is split, it becomes, '*adhika* + *ari*' by the pattern '*a* + *a* = *A*' from '*savarNadIrgha sandhi*'. '*adhika*' (means 'more') and '*ari*' (means 'enemy'). Both are root words and '*sandhi*' seems to be proper but the meaning yields 'more enemy', an incorrect translation. The primary requirement in translation is that the meaning of the context should not be disturbed.

VI. Special Cases Of '*ach sandhi*'

All the '*sandhis*' and the cases discussed above are related to single independent vowel. There are special cases in which either next or previous letters of the vowel is also to be checked in splitting. This ensures that the compound is formed by a particular '*sandhi*'.

For some '*sandhi*' rules, both the previous and next letters of the vowel are to be checked (TABLE XVIII). Following are the examples.

1) 'guNa sandhi': In specific cases, this 'sandhi' results in two letters instead of one in compound (TABLE VII). Sometimes more than one letter also to be checked since, to reduce time complexity in splitting i.e. six patterns causes to result in vowel 'a' but only three patterns can result 'ar'. Ex. For 'brahmarshi' – is a compound formed by two root words 'brahma' and 'Rshi'. All patterns are given in (TABLE XV, XVI).

TABLE XV. Possible Splitting By Observing Only Vowel '*a*'

Pattern	*sandhi*	Split forms	Result	NA/A
a + a	*akAra*	*brahma + arshi*	Fail	NA
u + a	*ukAra*	*brahmu + arshi*	Fail	NA
i + a	*ikAra*	*brahmi + arshi*	Fail	NA
aH + part2	*visarga*	*brahmaH + rshi*	Fail	NA
a + R	*guNa*	*brahma + Rrshi*	Fail	NA
A + R	*guNa*	*brahmA + Rrshi*	Fail	NA

TABLE XVI. Possible Splitting By Observing Two Letters '*ar*'

Pattern	*sandhi*	Split forms	Result	NA/A
a + R	guNa	brahma + Rshi	Succeeded	A
A + R	guNa	brahmA + Rshi	Fail	NA

From the Tables 15 and 16 it is observed that when a conjunction results in two or more letters, the total numbers of letters are to be observed for splitting.

2) 'yaNAdESa sandhi': Though 'yaNAdESa sandhi' is an 'ach sandhi', it results in generating a consonant in the compound along with the vowel.

Rule: When '*pUrva svara*' is '*i/u/R'* and '*para svara*' is '*i/u/R*' then, '*y*' replaces '*i', 'v*' replaces '*u*' and '*r*' + vowel replaces '*R*' as '*AdESam*' in the result (TABLE XVII).

Note: '*ya*, *va*, *ra*' are called '*yaNNs*' in Sanskrit grammar. When '*sandhi*' is formed, '*yaNNs*' comes as '*AdESam*'. That's why this '*sandhi'* is named '*yaNAdESa sandhi*'[3].

While checking vowels of the compound in vowel based splitting, if the vowel if preceded with the letter '*y/v/r*' then consider both the letters '*y/v/r'* + vowel to apply '*yaNAdESa sandhi*' rules in reverse engineering to find root words effectively.

TABLE XVII. All Patterns Of '*yaNAdESa sandhi*'

S.No	Pattern	Res	Example
1	*i + a*	*ya*	*ati + aMta = at*ya*Mta*
2	*i + A*	*yA*	*gWri + Arcana = gWr*yA*rcana*
3	*i + u*	*Yu*	*ati + unnati = at*yu*nnati*
4	*i + O*	*yO*	*dadhi + OdanaM = dadh*yO*danaM*
5	*u + a*	*va*	*madhu + annamu = madh*va*nnamu*
6	*u + A*	*vA*	*guru + AJna = gur*vA*Jna*
7	*R + A*	*rA*	*pirR + Arjitamu = pit*rA*rjitamu*

3. '*jastva sandhi*': This '*sandhi*' also results in specific consonants along with vowels in compound. These specific 'consonant + vowel' patterns are helpful (Except in some cases - refer Note of '*ukAra sandhi*'), in tracing exactly the root words by applying '*jastva sandhi*' rules.

Rule: when '*pUrva svara*' is 'k/c/T/t/p' and '*parasvara*' is a vowel/ *g/j/D/d/b/h/y/v/r*' then, '*g/j/D/d/b*' come as '*AdESam*' (TABLE XVIII).

TABLE XVIII. Vowel Patterns Of '*jastva sandhi*'

S.No	Pattern	Res	Example
1	*k + I*	*gI*	*vAk + ISuDu = vA*gI*SuDu*
2	*c + a*	*Ja*	*ac + aMtamu = a*ja*Mtamu*
3	*T + a*	*Da*	*shaT + aMgamu = sha*Da*Mgamu*
4	*t + A*	*dA*	*sat + AcAramu = sa*dA*cAramu*
5	*p + a*	*ba*	*kakup + aMtamu = kaku*ba*Mtamu*
6	*t + E*	*dE*	*tat + Ekamu = ta*dE*kamu*
7	*t + I*	*dI*	*jagat + ISuDu = jaga*dI*SuDu*

3) 'dvirukta TakAra sandhi': Occurrence of 'T' two times is called 'dvirukta TakAramu'.

Note: '*dvi*' means two, '*ukta*' means to tell, '*TakAramu*' means the letter '*Ta*'.

Rule: If the words '*kuru, ciru, kaDu, niDu, naDu*' connected with a vowel, then the letters '*ru', 'Du*' are replaced with '*TT*' (TABLE XIX).

TABLE XIX. All Patterns Of '*dvirukta TakAra sandhi*'

S.No	Pattern	Res	Example
1	*ru + e*	*TTe*	*ciru + eluka = ci*TTe*luka*
2	*ru + u*	*TTu*	*kuru + usuru = ku*TTu*suru*
3	*Du + a*	*TTa*	*kaDu + aluka = ka*TTa*luka*
4	*Du + i*	*TTi*	*naDu + illu = na*TTi*llu*
5	*Du + U*	*TTU*	*niDu + Urpu = ni*TTU*rpu*
6	*Du + A*	*TTA*	*kaDu + Ayata = ka*TTa*yata*
7	*Du + e*	*TTe*	*kaDu + edura = ka*TTe*dura*

While checking vowels of the compound in vowel based splitting, if the previous two letters of the vowel are '*TT*', then '*dvirukta TakAra sandhi*' rules are followed in reverse engineering to find out easily the root words.

4) 'visarga sandhi': This 'sandhi' also results in specific consonants along with vowels in some special cases. These 'consonant + vowel' patterns are helpful in tracing root words by applying 'visarga sandhi' rules.

Rule 1: when '*pUrva svara*' is '*H*' and '*para svara*' is '*S/sh/s*' then, '*S/sh/s*' come as '*AdESam*' respectively (TABLE XX).

TABLE XX. VOWEL PATTERNS OF '*VISARGA SANDHI*-1'

S.No	Pattern	Res	Example
1	(v)*H + S*	(v)*SS*	*tapaH + Sakti = tap**aSS**akti*
2	(v)*H + sh*	(v)*shsh*	*catuH + shashTi = cat**ushsh**ashTi*
3	(v)*H + s*	(v)*ss*	*manaH + sAkshi = man**ass**Akshi*

Note: 'v' stands for 'vowel'

Rule 2: when '*pUrva svara*' is '*H*', for '*para svara*' '*c/ch', 'AdESa*' is '*S*', for '*para svara*' '*T,Th*', '*AdESa*' is '*sh*', for '*para svara*' '*t,th*', '*AdESa*' is '*s*' respectively (TABLE XXI).

TABLE XXI. VOWEL PATTERNS OF '*VISARGA SANDHI*-2'

S.No	Pattern	Res	Example
1	(v)*H + c/ch*	(v)*sc/ch*	*manaH + calana= man**asc**alana*
2	(v)*H + T/Th*	(v)*shT/Th*	*ushaH + TaMka = ush**ash**TaMka*
3	(v)*H + t/th*	(v)*stth*	*manaH + tApam = man**ast**Apam*

Note: 'v' stands for 'vowel'

While checking vowels of the compound in this splitting, if next two letters of the vowel are '*SS/shsh/ss/sc/shT/st*', then applying '*visarga sandhi*' rules give better results in reverse engineering to find out the root words easily.

VII. DISCUSSIONS

a) Inflections:

Inflections are called '*vibhaktis*' which play an important role in Telugu grammar. In Telugu, inflections occur at the rear part of a word which leads in altering the original form of the root word. If the word is inflected then it is not possible to carryout splitting straightaway. All inflections must be separated and splitting is applied to obtain root words. Conjunction is possible not only with root words but also with inflected words.

For example, '*bhUmyAkASamutO*' is the compound which is inflected ('*tO*') at rear end. It is separated first and split rule is applied to obtain '*bhUmi' + 'AkASamu*' (TABLE XVII).

If '*pUrvapada*' is inflected and participated in conjunction, then it is difficult to find out root word. For example, in the sentence '*rAmuNNeMduku cUSAvu*' (why did you see Rama?) the compound is '*rAmuNNeMduku*' and is to be split. It is known that this is formed by two words '*rAmuNNi' + 'eMduku*'. Since the first word is inflected ('*rAmuDu' + 'ni'*) and such words are not available in database as they are. In such cases, splitting should be applied using morphology rules.

b) Plural forms:

Plural forms are very common in any language. If they are involved in conjunction, splitting becomes difficult. For example, '*kukkalarupulu*' (barking of dogs) is a compound formed by '*kukkala' + 'arupulu*'. '*pUrva pada*' is a plural term and is inflected. Formation of some plural words is not proper. For example '*baLLu*' can be the plural form of either '*baDi*' (school), or '*baMDi*' (a cart). But '*baDulu*' and '*baMDlu*' are the right plural forms of '*baDi*' and '*baMDi*' respectively.

But in normal conversations, corrupted form '*baLLu*' is intermittently used for representing plurals for both. Likewise, '*paLLu*' can also act as plural form for '*paMDu*' (fruit) and '*pannu*' (teeth). '*paMDlu*' and '*paLLu*' are the plural forms of the above respectively. When these are involved in conjunction, splitting becomes much difficult. For example '*baLLunnavi*' is the compound formed by either '*baDi + lu + unnavi*' (schools are there) or '*baMDi' + 'lu' + 'unnavi*' (carts are there).

c) Colloquial forms:

Colloquial or corrupted form of language is inevitable. These corrupted forms become impossible to split until unless they are maintained in Database. For example, '*rAmoDoccADu*' (Rama came) is the compound formed by '*rAmuDu' + 'occADu*' where '*occADu*' is the corrupted form of '*vaccADu*'. For successful splitting, either '*occADu*' must also be available in database or a rule must be made to morph/consider it as '*vaccADu*'.

d) Problems caused by conjunctions:

Some conjunctions create difficulties in identifying the root verb. For example, '*mEDipaMDujUDu*' (see the fig fruit) is the compound of root words '*mEDipaMDu' + ' cUDu*'. But according to conjunction rule ('*saraLAdESa sandhi'*) they must be '*mEDipaMDunu' + 'cUDu*' before conjunction. Here '*nu*' of the '*pUrva pada*' is removed and '*c*' of the '*parapada*' is converted to '*j*' and '*jUDu*' is not available in database.

'*gasaDadavAdESa sandhi*' also creates similar difficulties in splitting. For example, '*tallidaMDrulu*'(parents) is the compound formed by '*talli' + 'taMDrulu*'. According to conjunction rule, first letter of the '*parapada*' is converted from '*t*' to '*d*' and '*daMDrulu*' is not available in database. If '*da*' is morphed to '*ta*' for splitting, it leads to another difficulty. For example, '*Akalidappikalu*' (hunger and thirst) is the compound formed by '*Akali' + 'dappikalu*' by '*gasaDadavAdESa sandhi*'. If '*da*' of this compound is changed, then it becomes '*tappulu*' (mistakes) thus providing wrong translation.

VIII. CONCLUSION

Though there are many issues involved in splitting, splitting plays key role in MT. It paves a way to translate the source language as much as possible. Issues involved in the splitting can be solved by applying appropriate properly evolved morphological processes.

All possible patterns to observe in a compound for vowel based splitting given in TABLE XXII. Applying longest pattern as much as possible gives good results. Apply the rule of appropriate '*sandhi'* for splitting. When there are no multi-letter patterns available in compound, then it becomes mandatory to observe only single vowel for splitting. This may lead ambiguity in some cases. However, Vowel based splitting can separate at least one proper word from compound from left to right, if any. One can find more patterns for some special cases and can be included to split the compound very precisely.

TABLE XXII. TWO/THREE-LETTERS TO OBSERVE IN THIS SPLITTING

S.No	**V**	**If Nx / Pr**	**Pattern**	**Split pattern**	***Sandhi***
1	*a*	is – *r*	*Ar*	*a + R*	*guNa*
2	*a*	is – *r*	*ar*	*a + Ru*	*guNa*
3	*a*	is – *r*	*ar*	*A + R*	*guNa*
4	*a*	is – *r*	*ar*	*A + Ru*	*guNa*
5	*a*	is – *r*	*ar*	*aH+* part2	*visarga*
6	*i*	is – *r*	*ir*	*iH* + part2	*visarga*
7	*I*	is – *r*	*Ir*	*IH* + part2	*visarga*
8	*u*	is – *r*	*ur*	*uH+* part2	*visarga*
9	V	is – *y*	*y+* V	*i* + vowel	*yaNAdESa*
10	*a*	is – *v*	*va*	*u + a*	*yaNAdESa*
11	*A*	is – *v*	*vA*	*u + A*	*yaNAdESa*
12	*A*	is – *v*	*vA*	*U + A*	*yaNAdESa*
13	*a*	is – *r*	*ra*	*R + a*	*yaNAdESa*
14	V	is – *g*	*g+* V	*k* + V	*Jastva*
15	V	is – *j*	*j* + V	*c* + V	*Jastva*
16	V	is – *D*	*D* + V	*T* + V	*Jastva*
17	V	is – *d*	*d+* V	*t* + V	*Jastva*
18	V	is – *b*	*b+* V	*p* + V	*Jastva*
19	V	is – *TT*	*TT* + V	*Du/ru* + V	*dviruktTakAr*
20	V	is – *SS*	V + *SS*	V+*H + S*	*visarga*
21	V	is – *shsh*	V + *shsh*	V+*H+sh*	*Visarga*
22	V	is – *ss*	V + *ss*	V+*H+s*	*visarga*
23	V	is – *sc*	V + *sc*	V+*H + c*	*visarga*
24	V	is – *sch*	V + *sch*	V+*H + ch*	*visarga*
25	V	is – *shT*	V + *shT*	V+*H + T*	*visarga*
26	V	is – *shTh*	V + *shTh*	V+*H + Th*	*visarga*
27	V	is – *st*	V + *st*	V+*H+t*	*visarga*
28	V	is – *sth*	V + *sth*	V+*H+th*	*visarga*

* Here ‘pr’ for previous, ‘nx’ for next, and ‘V’ for vowel.

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] Akshar Bharati, Amba Kulkarni and V Sheeba, “Building a Wide Coverage Sanskrit Morphological Analyzer: A Practical Approach”, The First Nat. Symp. on Modelling and Shallow Parsing of Indian Languages, IIT Bombay, 2006.

[7] Malhar Kulkarni, Chaitali Dangarikar, Iravati Kulkarni, Abhishek Nanda, and Pushpak Bhattacharya, “Introducing Sanskrit Wordnet”, Proc. of Global Wordnet Conf. 2010, Mumbai, India, 2010.

[8] Arthur. A. McDonell, “Sanskrit Grammar for Students”, 3rd Edition, Oxford University Press, 1926.

[9] Joshi Shripad S. “Sandhi Splitting of Marathi Compound Words”, Int. J. on Adv. Computer Theory and Engg., Vol. 2 Issue 2, 2012

[10] S. Varakhedi, V. Jaddipal and V.Sheeba, “An Effort To Develop A Tagged Lexical Resource For Sanskrit”, FISSCL Paris, Oct 2007.

[11] Anil Kumar, Vipul Mittal, Amba Kulkarni “Sanskrit Compound Processor”, Sanskrit Computational Linguistics, pp. 57-69, 2010

[12] Shathaka Sagaram, available at http://shathakasagaram.blogspot.in/2011/05/blog-post_03.html

[13] Jasti Suryanarayana, “Sanskrit for Telugu Students” , Sri Balaji Printers, Tirupati, 1993

[14] Divakrala Venkata Avadhani, “Telugu in Thirty Days”, Andhra Pradesh Sahithya Academy, 1976.

[15] Akshar Bharati, Vineet Chaitanya, Rajeev Sangal, “Natural Language Processing – A Paninian Perspective”, Prentice-Hall of India, 1994.

[16] T. Suryakanthi, Dr. S. V. A. V. Prasad and Dr. T. V. Prasad “Translation of Pronominal Anaphora from English to Telugu Language”, Int. J. of Adv. Computer Sc. App., Vol. 4 Issue 4, 2013.

APPENDIX

Pronunciations: The letters should be pronounced normally as in English, except when they are italicized. If so, follow the TABLE XXIII.

TABLE XXIII. ROMAN TELUGU 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 - jack	*Ru*	roo in – roof	*O*	oa in - Roar
i	i in - His	*e*	e in - When	*W*	Ou in - out
I	Ea in- east	*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 stands for Roman Telugu and the capitalized letters should be pronounced with greater emphasis on them.