Solving Ambiguities in Indonesian Words by Morphological Analysis Using Minimum Connectivity Cost*

PRESENTED BY ELVIRA NURFADHILAH

*Based On Work By Mohammad Teduh Uliniansyah, Shun Ishizaki, And Kiyoko Uchiyama

Curriculum Vitae

Name	**: Elvira Nurfadhilah**
Working at	**: Agency for the Assessment & Application of Technology (BPPT)**
Laboratory	**: Intelligence Computing Laboratory (ICL)**
Specialisation field	**: Image Processing and Natural Language Processing**
Email	**: elvira.nurfadhilah@bppt.go.id/ elvira.nurfadhilah@gmail.com /**
Educational background	
Under Graduate	**: Bogor Agriculture University in Computer Science (2011)**
Graduate	**: Bogor Agriculture University in Computer Science (2015)**
Joined at BPPT	**: 2014**

ICL
Intelligent Computing Laboratory

Intelligent Computing Laboratory (ICL) at the Center for Information and Communication Technology (PTIK), BPPT.

ICL deals with *image processing, computer vision, language technology and signal processing.*

Natural Language Processing	Biometric	ICT Health
❑ Portal Bahasa (Stemmer and Concordance) ❑ Statistical Machine Translation ❑ Text-To- Speech ❑ Etc.	❑ Fingerprint and Latten Fingerprint ❑ Iris ❑ Face and Face sketch ❑ Blood vessel ❑ etc.	❑ Developing a malaria diagnosis tool based on images of thin and thick smears.
Badan Pengembangan dan Pembinaan Bahasa Mitra Eksternal	Dukcapil (Kemendagri) dan Pemerintah Daerah Mitra Eksternal	Lembaga Biologi Molekuler Eijkman Mitra Eksternal

Outline

- Background
- Experimental Data
- Method
- Results and Discussion
- Utilizing the proposed technique for Wordnet
- Demo Program

BACKGROUND

Ambiguities

Ambiguities arise when a single lexical word may have been created by more than one pos-sible combination of affixes.

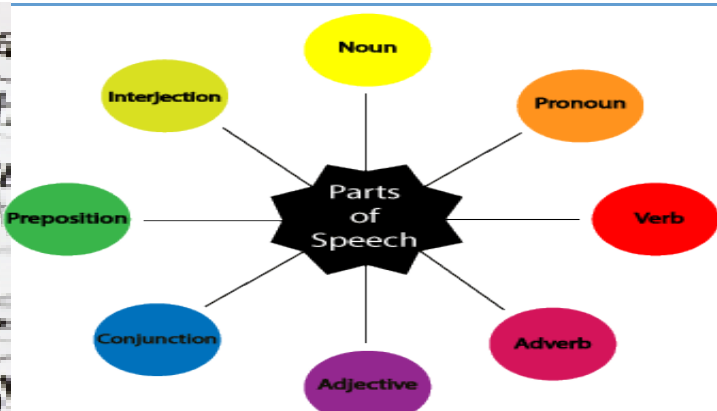
Example:

beruang:

- **beruang (Noun Animal)**
- **ber (uang (Noun Concrete)) : Verb Intransitive**
- **be (ruang (Noun Abstract Concept)) :**

  Verb Intransitive

EXPERIMENTAL DATA



Corpus Data

A corpus consists of articles (politics,economics,sports,etc.) downloaded from"Kompas"daily newspaper website (http://www.kompas.com). The corpus contains 20,579,771 words in 1,105,156 sentences.

Rule for Affixes

First Field	Second Field
[meng;kan;k]	(IDK,^keluar:IDK,IDKT;IDB,^kurai,^karah,^kait,^kisi, ^koperasi,^korban:IDK,IDKT;IDB,IDBAU,korban:IDK,IDKT;IDB,IDBKU,kait:IDK,IDKT;IDS,^kosong:IDK,IDKT;IDS,IDSADA,kosong:IDK,IDKT;IDT,^kembali:IDK,IDKT;IDD:IDK,IDKT)
[mem;lah;]	(IDK:IDK,IDKT;IDB:IDK,IDKT;IDS:IDK,IDKT)
[me;kan;]	(IDK:IDK,IDKT;IDK,mati:IDS,IDSEVA;IDB,^rupa,^madu:IDK,IDKT;IDB,rupa:IDK,IDKH;IDS,^lahir:IDK,IDKT;IDD:IDB,IDBKU,IDBAGE;IDT:IDK,IDKT;IDT:IDK,IDKT)
[me;kanlah;]	(IDK:IDK,IDKT;IDB:IDK,IDKT;IDS:IDK,IDKT;IDD:IDB,IDBKU,IDBAGE;IDT:IDK,IDKT)
[me;alkan;]	(IDB:IDK,IDKT)

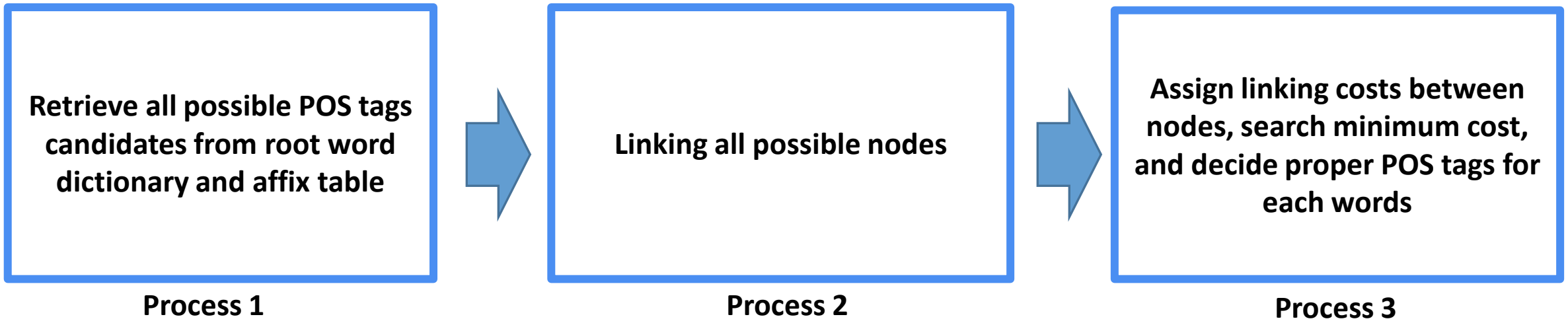
There are more than 800 combinations of affixes (prefixes, suffixes, and infixes)

List of Part of Speech Used (2)

Tag	Part of Speech	Tag	Part of Speech
IDBMON	Noun Abstract Money	IDBBLD	Noun Building
IDBNTR	Noun Abstract Title	IDBANM	Noun Animal
IDBKEJ	Noun Abstract Event	IDBKU	Noun Concrete
IDBORG	Noun Abstract Organization	IDBBRA	Plural
IDBWKT	Noun Abstract Time	IDBPLC	Name Place
IDBSCI	Noun Abstract Science	IDBB	Name
IDBSOS	Noun Abstract Art	IDSADA	Adjective Condition
IDBS	Noun Abstract Unit	IDSWRN	Adjective Color
IDBLOK	Noun Abstract Location	IDSUKR	Adjective Quantitative
IDBAKS	Noun Abstract Action	IDSEVA	Adjective Judgement
IDBKON	Noun Abstract Concept	IDSFEL	Adjective Feeling
IDBPRS	Noun Abstract Process	IDSIDR	Adjective Sense
IDBMED	Noun Abstract Medical	IDSBTK	Adjective Form
IDBAU	Noun Abstract	IDSWKT	Adjective Time

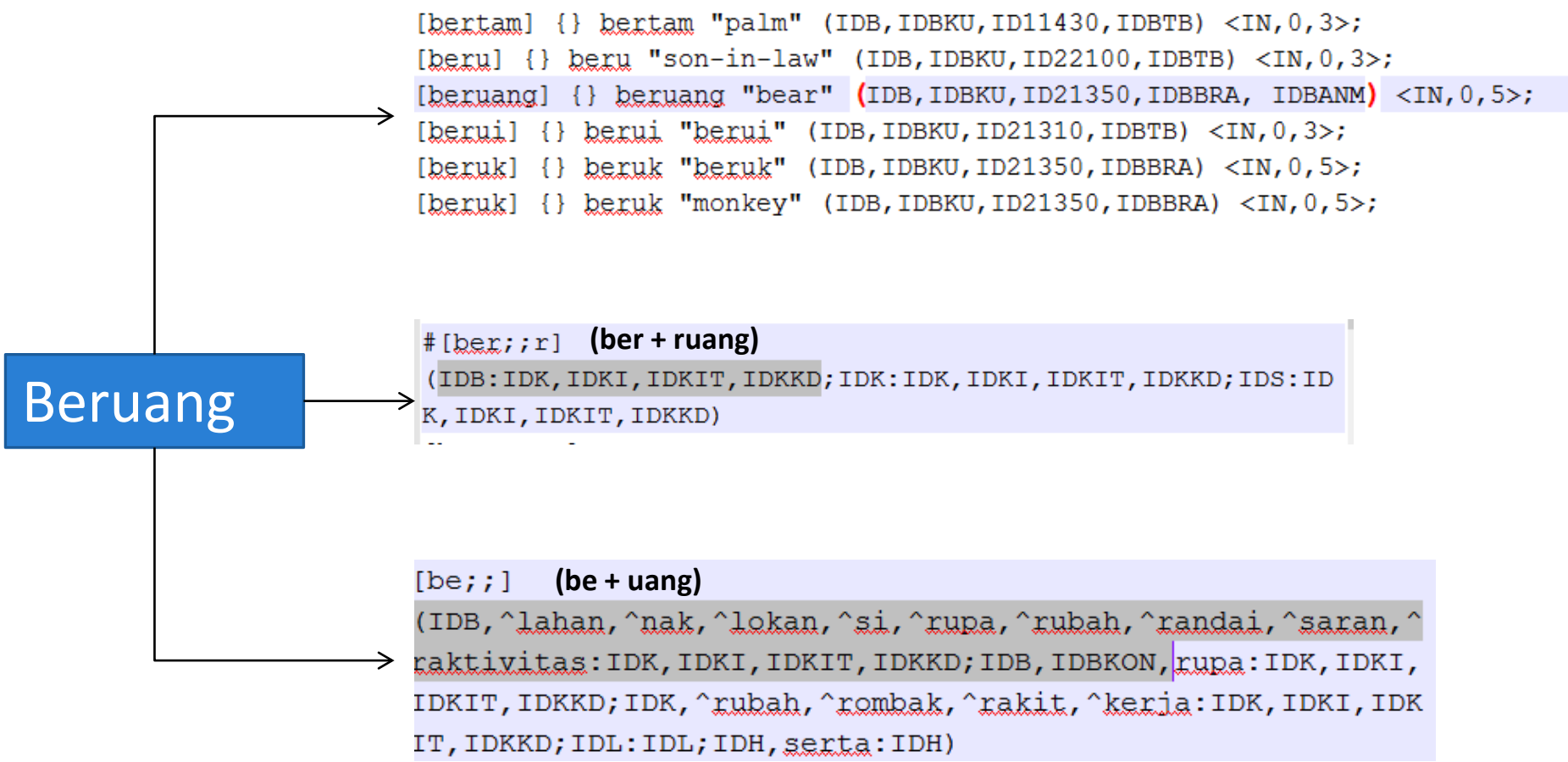
METHOD

Flow of the Morphological Analysis Process



PROCESS 1

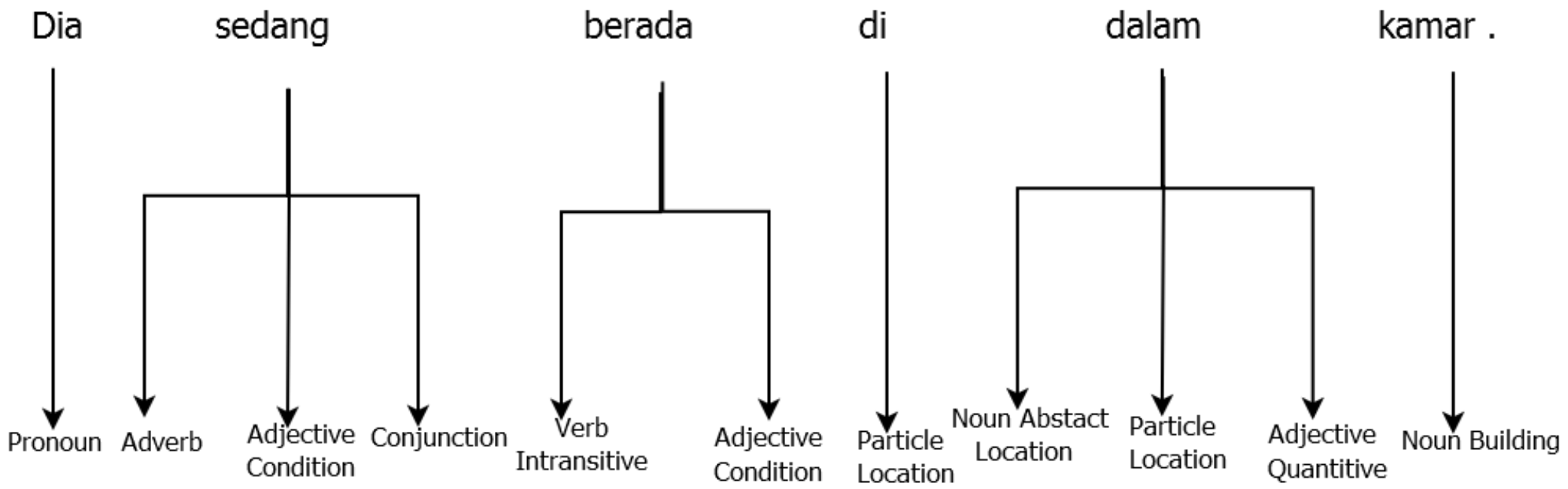
(Finding all possible tags in one word)



PROCESS 1

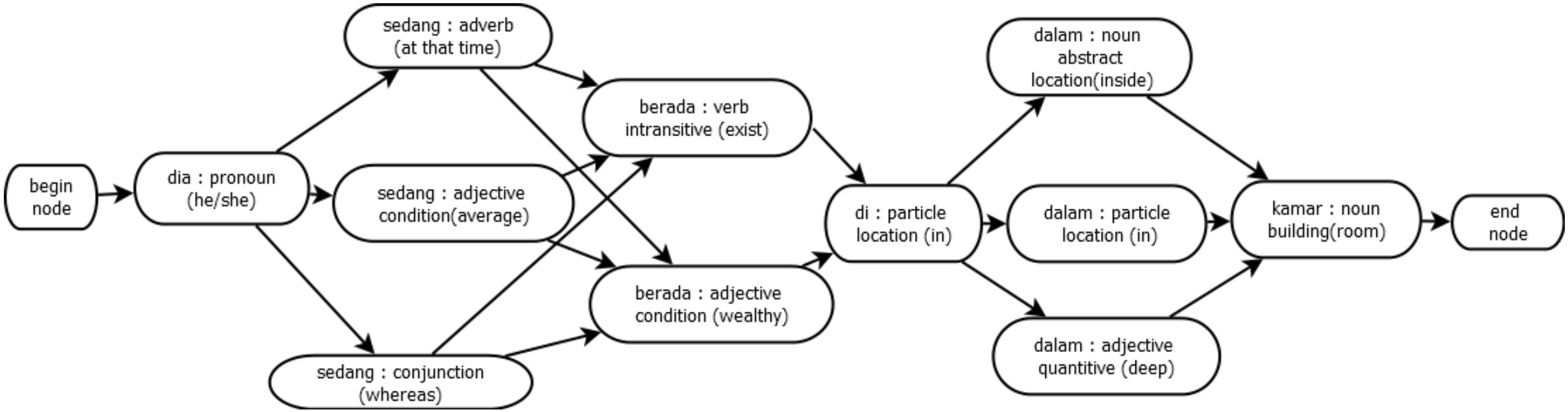
(Finding all possible words tag in one sentence)

Example



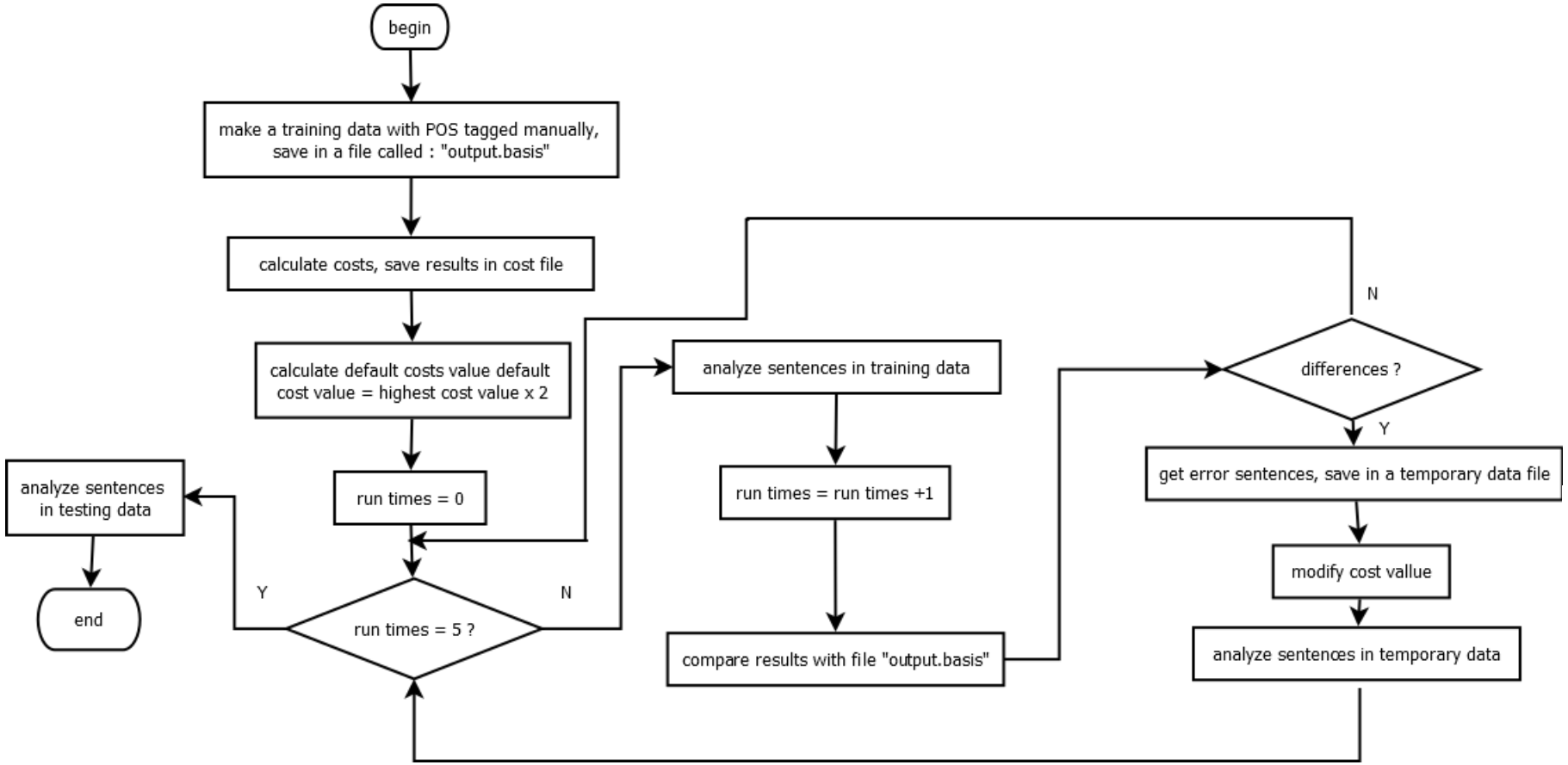
PROCESS 2

(Linking all possible Nodes)



PROCESS 3

(Training process to get optimum linking costs)



$If\ \Delta = 1,\ do\ nothing\ to\ \Delta$

$Else\ if\ 0.125 < \Delta < 1,\ \Delta = \Delta * 0.5$

$If\ the\ cost\ correct:$

$Cost = Cost + \Delta$

$Else:$

$Cost = Cost - \Delta$

PROCESS 3
(Calculating cost of links between nodes)

Let p1 be one POS and p2 another one, where p2 directly follows *pi.* The cost of the pair (p1,p2) is:

$$Cost\ (p1,p2) = {}^{2}log(N/n(p1,p2))$$

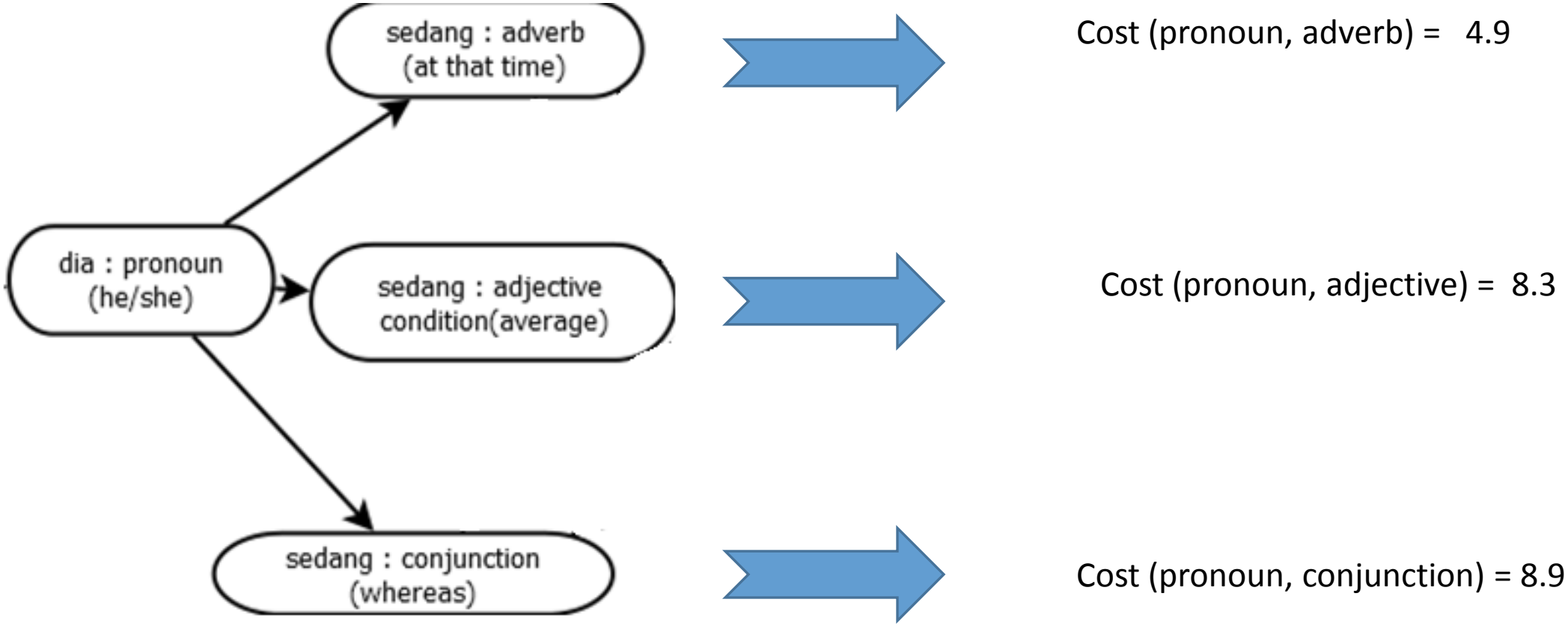
where n(p1,p2) is the number of (p1,p2) pairs which appear in the data. N is the total number of all of the pairs of POS tags in the data.

List of Linking Costs Between Nodes

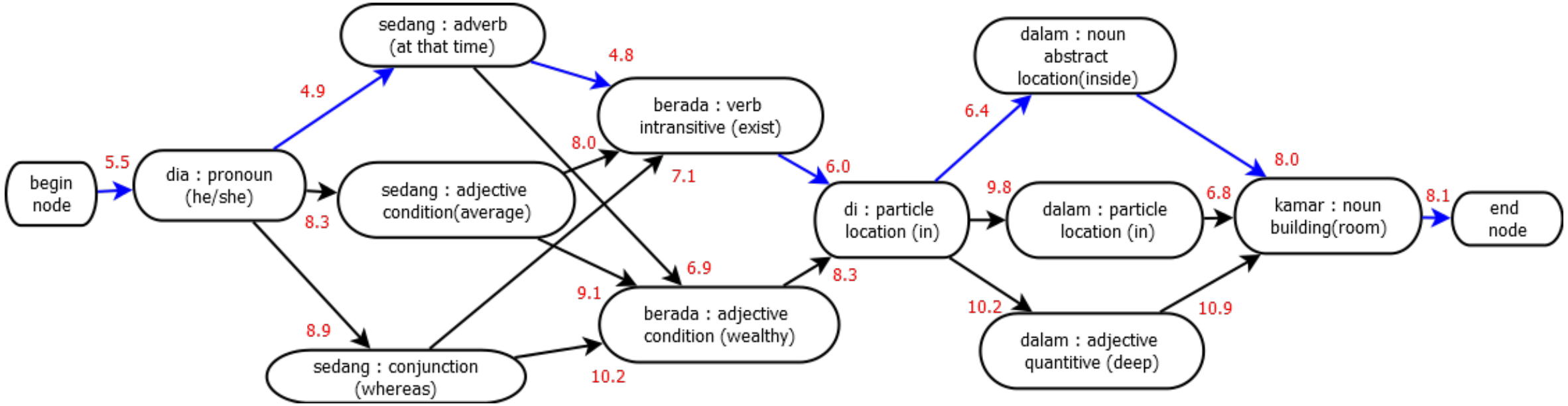
Example :

```
[ |Adjective,Comma] : 10.1778563959258
[ |Adjective,Dot] : 11.5641507570457
[ |Adjective,Possesion3rd,Comma] : 11.5641507570457
[ |Adjective,Possesion3rd] : 11.5641507570457
[ |AdjectiveCondition,Comma] : 11.5641507570457
[ |AdjectiveCondition,Possesion3rd,Comma] : 11.5641507570457
[ |AdjectiveCondition,Possesion3rd] : 11.5641507570457
[ |AdjectiveCondition] : 8.73093741298946
[ |AdjectiveJudgement,Comma] : 11.5641507570457
[ |AdjectiveJudgement,Possesion3rd,Comma] : 9.48470921536584
[ |AdjectiveJudgement,Possesion3rd] : 10.4655384683776
[ |AdjectiveJudgement] : 7.98063181858956
[ |AdjectiveQuantitative,Possesion3rd,Comma] : 11.5641507570457
[ |AdjectiveQuantitative,Possesion3rd] : 9.77239128781762
[ |AdjectiveQuantitative] : 9.1662554842473
[ |AdjectiveSense,Possesion3rd] : 11.5641507570457
```

Illustration of Linking Nodes



PROCESS 3
(Search minimum cost)



An Example of Possible Analysis for a Simple Input Sentence

RESULTS & DISCUSSION

Training and Testing Data

	Number of Words	Number of Sentences
Training Data A	53,432	3,205
Training Data B	98,388	5,977
Test Data 1	4,994	299
Test Data 2	4,998	319
Test Data 3	5,005	292
Test Data 4	5,007	300
Test Data 5	5,017	312

Training Results

	Training Data A		Training Data B	
Times of training	Errors	Accuracy	Errors	Accuracy
1	1062 (847)	98.01 %	2007 (1600)	97.96 %
2	447 (391)	99.16 %	1084 (947)	98.89 %
3	408 (358)	99.23 %	1148 (985)	98.83 %
4	363 (325)	99.32 %	974 (817)	99.01 %
5	415 (364)	99.22 %	1059 (900)	98.92 %

Test Results Using Worst Cost Data

	Training data A			Training data B		
	Errors	Accuracy	Score	Errors	Accuracy	Score
Test Data 1	136 (96)	97.28 %	9.73	134 (93)	97.32 %	9.73
Test Data 2	114 (85)	97.72 %	9.78	102 (79)	97.96 %	9.8
Test Data 3	106 (77)	97.88 %	9.79	104 (77)	97.92 %	9.79
Test Data 4	144 (102)	97.12 %	9.71	143 (103)	97.14 %	9.71
Test Data 5	146 (109)	97.09 %	9.71	136 (106)	97.29 %	9.73
Average	129 (94)	97.42 %	9.74	124 (92)	97.53 %	9.75

Test Results Using Best Cost Data

	Training data A			Training data B		
	Errors	Accuracy	Score	Errors	Accuracy	Score
Test Data 1	88 (70)	98.24 %	9.82	85 (66)	98.30 %	9.83
Test Data 2	90 (71)	98.2 %	9.82	67 (56)	98.66 %	9.87
Test Data 3	69 (59)	98.62 %	9.86	65 (58)	98.70 %	9.87
Test Data 4	105 (81)	97.90 %	9.79	80 (61)	98.40 %	9.84
Test Data 5	104 (84)	97.93 %	9.79	86 (74)	98.29 %	9.83
Average	91 (73)	98.18 %	9.82	77 (63)	98.47 %	9.85

Reference

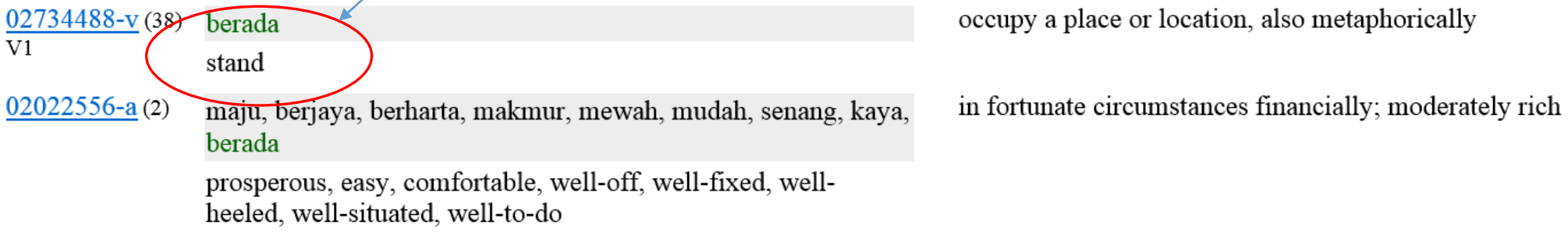
- Uliniansyah MT, Ishizaki S, Uchiyama K. 2004. **Solving Ambiguities in Indonesian Words by Morphological Analysis Using Minimum Connectivity Cost. Journal of Natural Language Processing, Vol. 11, No. 1**
- Kridalaksana, H.(1996). **Pembentukan Kata dalam Bahasa Indonesia. PT Gramedia Pustaka Utama.**

Utilizing the proposed technique for WordNet

We can use Wordnet and this technique to choose the proper sense of the word in a sentence.

Contoh :

Dia \pronoun sedang\adverb berada\verb intransitive di\partical location dalam\noun abstract location kamar\noun building



02734488-v (38) V1	berada stand	occupy a place or location, also metaphorically
02022556-a (2)	maju, berjaya, berharta, makmur, mewah, mudah, senang, kaya, berada prosperous, easy, comfortable, well-off, well-fixed, well-heeled, well-situated, well-to-do	in fortunate circumstances financially; moderately rich

DEMO PROGRAM


Terimakasih
Aguyje
Nani
Modupe
Shukria
Nuhun
Dank Je
Talofa
Dhannyabad
Dank
Kiitos
Merci beaucoup
Blagodaria
Akpe
Arigato Gozaimas
Mahalo
Tingki
Danke
Kam ouen
Thoinks
Grazzi
Gracias
Haika
Toda
Tatenda
Syukriya
Thank You
Mamnuun
Aabar
Danki
Dank u
Mese
Obrigado
Tusen takk
Tangur
Barakallahu fik
Syukron
Rahmet sizge
Grazie
Salamot
Spasiba
Thenk Ye
Gamsa-hamnida
Tayu'an
Vinaka
Shukram
Shukria
Xie-xie
Hvala
Nandi
Tanggio
Merci
Dankie
Matur Nuwun
Marahaba
Tenkiu
Salamat po
Muchas gracias
Maururu
Bhala Hove
Tsikomo
Mese
Webale
Mahad sanid
Ahsante
Dyakuyu
Ngiyabonga

Demo Program

```
elvira@elvira-BPPT:~/paperdanfile2analisamorfologi$ perl ./morfo.pl
Processing word dictionary
Processing Affix Tables
Processing frequency table

Type sentence: Dia sedang berada di dalam kamar
TMPWORD: sedang: , TMPREST: AdjectiveCondition|Adverb|Conjunction
TMPWORD: berada: , TMPREST: ber(ada(VerbIntransitive)): VerbIntransitive|ber(ada(VerbIntransitive)): AdjectiveCondition
TMPWORD: dalam: , TMPREST: AdjectiveQuantitative|PartikelLocation
Dia: Pronoun^
  sedang: Adverb
  berada: VerbIntransitive
di: PartikelLocation
  dalam: PartikelLocation
kamar: NounBuilding#
```