Average Referential Distance

Internal Report
Erwin R. Komen
August 29th, 2011

1 Introduction

Syntactically annotated corpora that have been enriched with coreferential information allow for the calculation of the "Average Referential Distance" (Givón, 1983). Givón defines the "Referential Distance" as the amount of clauses between a Noun Phrase (NP) and its antecedent. The "Average Referential Distance", which I will refer to as "ARD" from now on, calculates the average of all referential distances for the NPs in a particular construction or word order. Givón was able to differentiate languages on the basis of ARDs measured for constructions like left dislocation and right dislocation.

I propose looking at several different constructions and word orders that we expect to change with respect to the amount of clauses their NPs will have their antecedents. We have been talking about the area before the finite verb (or auxiliary) in the main clause, for instance, suggesting that non-subject NPs in this area refer back more locally in Old English (OE) than they do in present-day English (PDE). So we would expect to see an increase in the ARD value of the non-subject NPs in the preverbal area in general.

Other positions that could be of interest are: (a) left dislocated NPs, (b) the NPs within preverbal PPs, and (c) the preverbal subjects themselves.

2 Data

The CorpusStudio has been extended with a function `ru:ard($node, $strType)`. This function, operating only within Xquery on `psdx` files, adds the referential distance of the constituent `$node` to the collection of those with the construction named `$strType`. The CorpusStudio project called "AvRefDist_V1" makes use of this function, and calculates the ARD for the constructions mentioned above. The results are presented in Table 1.

Type	**OE**			**ME**			**eModE**			**LmodE**		
	Count	ARD	Sdev	Count	ARD	Sdev	Count	ARD	Sdev	Count	ARD	Sdev
ldnp (all)	*16*	1,25	2,63	*1*	0,00	0,00	*0*	0,00	0,00	*0*	0,00	0,00
ldnp (>0)	*4*	5,00	3,00	*0*	0,00	0,00	*0*	0,00	0,00	*0*	0,00	0,00
np (all)	*77*	2,69	8,00	*13*	1,38	2,37	*9*	0,22	0,42	*10*	10,60	26,89
np (>0)	*43*	4,81	10,22	*7*	2,57	2,72	*2*	1,00	0,00	*5*	21,20	34,94
pp (all)	*21*	2,62	6,77	*5*	17,20	21,43	*10*	1,30	2,00	*62*	1,32	3,03
pp (>0)	*10*	5,60	8,90	*5*	17,20	21,43	*4*	3,25	1,92	*21*	3,90	4,13
sbj (all)	*294*	6,00	50,94	*205*	5,61	10,88	*213*	4,36	11,38	*1028*	5,18	15,59
sbj (>0)	*251*	7,03	55,07	*179*	6,43	11,41	*146*	6,38	13,26	*772*	6,91	17,66

Table 1 Average referential distance for NPs from four constructions

The first rows, labelled "ldnp", list the ARD for left dislocated NPs within main clauses. Those are apparently so rare in the available data, that there are no occurrances in eModE and LmodE.

The next two rows are labelled "np", and they contain the ARD values for the non-subject NPs occurring before the finite verb or auxiliary in the main clause. The first of the

two “np” rows contains the measure for *all* NPs, whether they actually have an antecedent or not. The second of these two “np” rows bears the addition “>0”, which means that it only counts referential distances that are larger than “0”. This means that NPs that are labelled “New” and “Assumed” are not incorporated in the ARD measure.

There seems to be a downward trend from OE to ME into eModE, but the LmodE data show completely different results. However, the standard deviation values reveal that there just is too little data available to provide statistically significant answers. They could, alternatively, reveal that it just makes no sense to speak of an ARD at all.

The NPs within preverbal PPs are in the rows labelled “pp”. Their numbers are too low to gain any significance.

The preverbal subjects are in the rows labelled “sbj”, and they offer some interesting data. Their numbers are huge enough to calculate good averages. However, the standard deviations for these averages are so high, that we can, perhaps, conclude that the ARD measure is not revealing enough for any construction.

3 Discussion

Measurement of the “Average Referential Distance” as defined by Givón has been made possible with the coreferentially enriched syntactically annotated texts from English historical corpora. I have used this measure on a number of different constructions that we thought would change in time, from the point of view of referentiality. Statistical evaluation of the data indicate that the “Average Referential Distance” may not be tied with enough significance to the constructions I looked at.

4 References

Givón, Talmy. 1983. “Topic continuity in discourse: an introduction”. *Topic continuity in discourse: a quantitative cross-language study*, ed. by Talmy Givón. Amsterdam, John Benjamins.

5 Appendix: Corpus Research Project

This appendix gives the main query of the CorpusResearchProject called “AvRefDist_V1” that has been used to measure the averager referential distances.

5.1 AvRefDist_V1

The main query of this project is called “matPreVb-S.xq”, and is listed here.

```
{
 for $search in //eTree[tb:HasLabel(@Label, $_matrixIP)]
    (: Get the finite verb :)
    let $vb  := ru:one($search, 'FirstChild', $_finiteverb)

    (: Get all the NPs preceding the finite verb that qualify :)
    let $np  := $vb/preceding-sibling::eTree[tb:Like(@Label, $_anynp)         and
                                             not(tb:Like(@Label, $_noobject)) and
                                             not(child::eLeaf/@Type = 'Star') ]
    (: Select the subject from it :)
    let $sbj     := $np[tb:Like(@Label, $_subject)][last()]

    (: See if there is another NP left :)
    let $oth     := if (exists($sbj)) then $np[not(@Id = $sbj/@Id)][1] else ()

    (: See if there is a PP preceding the subject :)
    let $ppall   := $sbj/preceding-sibling::eTree[tb:Like(@Label, $_anypp)   and
                          empty(child::eTree[tb:Like(@Label, $_anyClause)]) and
                          exists(child::eTree[@Label = 'P'])                and
                          exists(child::eTree[tb:Like(@Label, $_anynp)]) ]
    (: Get the NP dominated by the FIRST PP :)
    let $ppnp    := $ppall[1]/child::eTree[tb:Like(@Label, $_anynp)]

    (: See if there is a left dislocated NP :)
    let $ldnp    := ru:one($search, 'iDoms', $_anyNpLfd)

    (: Note the referential distance of the subject :)
    let $sbjOut := ru:ard($sbj, 'sbj')

    (: Note the referential distance of the other NP, if exists :)
    let $othOut := if (exists($oth)) then
                     ru:ard($oth,'np')
                   else
                     true()

    (: Note the referential distance of the NP inside a PP, if exists :)
    let $ppOut := if (exists($ppnp)) then
                     ru:ard($ppnp,'pp')
                   else
                     true()

    (: Note the referential distance of LFD, if it exists :)
    let $lfdOut := if (exists($ldnp)) then ru:ard($ldnp, 'ldnp') else true()

where (
         exists($vb)          and
         exists($sbj)         and
         not($sbjDist='')     and
         $sbjOut              and
         $othOut              and
         $ppOut               and
         $lfdOut
      )

 (: Output providing a message in $msg :)
 return tb:MyForestMsg($search, $msg)
}
```