
Manuel González González (2015): "O preprocesado lingüístico nos programas de síntese de voz", en Francisco Dubert García / Gabriel Rei-Doval / Xulio Sousa (eds.): *En memoria de tanto miragre. Estudos dedicados ó profesor David Mackenzie.* Santiago de Compostela: Servizo de Publicacións da Universidade de Santiago de Compostela, 91-101.

Instituto da Lingua Galega

En memoria de tanto miragre

Estudos dedicados ó profesor David Mackenzie

Edición ó coidado de

Francisco Dubert García
Gabriel Rei-Doval
Xulio Sousa

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En memoria de tanto miragre

INSTITUTO DA LINGUA GALEGA

En memoria de tanto miragre
Estudos dedicados ó profesor David Mackenzie

Edición ó coidado de
FRANCISCO DUBERT GARCÍA
GABRIEL REI-DOVAL
XULIO SOUSA

2015
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En memoria de tanto miragre : estudos dedicados ó profesor David Mackenzie / edición ó coidado de Francisco Dubert García, Gabriel Rei-Doval, Xulio Sousa. – Santiago de Compostela : Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico, 2015. – 261 p. ; 24 cm.

Precede ó tít.: Instituto da Lingua Galega

D.L. C 1193-2015. – ISBN: 978-84-16183-96-8

1. Mackenzie, David – Crítica e interpretación. 2. Galego (Lingua). 3. Galego-portugués (Lingua)—Antes de 1500. 4. Literatura galega. 5. Literatura española. 6. Lingüística histórica iberorrománica. I. Dubert García, Francisco, ed. lit. II. Rei-Doval, Gabriel, ed. lit. III. Sousa, Xulio, ed. lit. IV. Universidade de Santiago de Compostela. Servizo de Publicacións e Intercambio Científico, ed. V. Instituto da Lingua Galega

801 Mackenzie, David
806.99
806.90/99 "04/14"
869.9

Este libro publícase coa axuda financeira da Secretaría Xeral de Universidades (Xunta de Galicia) ó grupo de investigación *Filoloxía e lingüística galega* (GI-1743), da Universidade de Santiago de Compostela.

Deseño de cuberta
Raquel Vila Amado

Maquetación
Raquel Vila Amado

Imprime
Imprenta Universitaria
Campus Vida
15782 Santiago de Compostela

Edita
Servizo de Publicacións
Campus Vida
15782 Santiago de Compostela
usc.es/publicacions

Dep. Legal C 1193-2015

ISBN 978-84-16183-96-8

En memoria de tanto miragre
Estudos dedicados ó profesor David Mackenzie
Universidade de Santiago de Compostela, 2015, ISBN 978-84-16183-96-8, pp. 91-101

O preprocesado lingüístico nos programas de síntese de voz

MANUEL GONZÁLEZ GONZÁLEZ
Instituto da Lingua Galega - Universidade de Santiago de Compostela

Podemos dicir que nos últimos anos as tecnoloxías da fala están de moda, e están de moda porque nos facilitan o traballo, permiten unha maior eficiencia e posibilitan logros que eran practicamente impensables hai ben poucos anos. Xa dificilmente se entende unha sociedade moderna sen unha forte presenza das tecnoloxías da fala.

Hoxe cando telefonamos ao Hospital ou á Universidade a nosa chamada non é recibida por un telefonista, senón por unha máquina que nos pregunta se coñecemos o número do cuarto ou a extensión do gabinete cos que nos queremos comunicar, e o propio sistema vainos dirixindo ata establecer a comunicación que desexamos. Se queremos obter por teléfono a información meteorolóxica dun determinado lugar, o máis normal será que nos responda unha máquina que atenda a nosa demanda e responda a nosa solicitude. Na actualidade é normal que un invidente poida consultar unha páxina web e poida navegar por ela sen dificultade ou que poida consultar un dicionario en liña facendo uso da voz e do oído. Son moitos xa os que utilizan un ditáfono para escribir un texto, en vez de introducilo a través do teclado. Todo isto é posible grazas ás tecnoloxías da fala, fundamentais para a automatización do mundo da comunicación e da información e para o acceso a determinados servizos por parte de persoas con algunha minusvalía ás que estes servizos lles estaban totalmente vedados non hai moito tempo.

Dentro das tecnoloxías da fala, a síntese de voz ocupa un lugar de especial relevancia, xa que permite a conversión de calquera texto escrito nun texto oral sen a intervención humana directa. Hai diversos procedementos para lograr a voz sintetizada, pero hoxe os sistemas máis eficientes son os que se basean na concatenación de unidades.

Para a lingua galega construíronse ata o momento tres conversores texto--voz: a) *Cotovía*, desenvolvido no Centro Ramón Piñeiro para a investigación en humanidades por enxeñeiros de telecomunicacións da Universidade de Vigo (baixo a dirección de Carme García Mateo) e por lingüistas da Universidade de Santiago (baixo a dirección de Manuel González González), b) o elaborado por Telefónica I+D, fundamentalmente para uso da propia empresa; e c) o de Loquendo, a corporación multinacional de tecnoloxía de software, con sede central en Torino, líder en tecnoloxías da fala. Os tres sistemas en funcionamento ata o momento constan de dous grandes módulos: un módulo lingüístico e un módulo acústico, e nos tres participei en maior ou menor medida na elaboración e posta a punto dos seus respectivos módulos lingüísticos.

Que é o preprocesado lingüístico

Existe unha fase previa, habitualmente coñecida como de *preprocesado lingüístico* ou de *normalización do texto*, que é en certo modo marxinal por ser anterior ao momento en que entran en pleno funcionamento as distintas análises lingüísticas que proporcionarán a información necesaria para unha transcrición fonética e información prosódica adecuadas. Pero o feito de ser unha fase periférica non quere dicir que non sexa fundamental para o correcto funcionamento do módulo lingüístico e do módulo acústico, e que estea exenta de dificultades e de problemas. A súa resolución dunha maneira máis satisfactoria ou menos satisfactoria condicionará en gran medida o resultado dos procesos posteriores.

Nesta fase previa lévase a cabo unha normalización do texto, proceso durante o que é necesario interpretar e expandir as secuencias gráficas que non corresponden a unha palabra ortográfica pronunciable, como son os signos de puntuación con valor non lingüístico, os números arábigos ou romanos, os símbolos, as siglas ou as abreviaturas.

É unha fase problemática á que non se lle adoita prestar a suficiente atención, con dificultades ás veces difíciles de salvar, debido á multiplicidade de valores que estes signos posúen no uso lingüístico, ás distintas lecturas que estes poden ter, a problemas de concordancia de xénero na lectura dos números que presentan moción xenérica (*vinte e un homes/vinte e unha mulleres, dous cabalos/dúas eguas),* ao duplo, triplo ou mesmo cuádruplo valor que poden ter algunhas siglas ou algunhas abreviaturas etc.

Podemos dicir que a maioría das dificultades que poden aparecer na fase de preprocesado lingüístico veñen dadas principalmente por dous factores: a) a posibilidade de dobre, tripla ou cuádrupla lectura que pode ter un determinado elemento que cómpre desenvolver); b) a falta de estandarización na utilización de determinados signos e de determinados usos.

Vexamos algunhas das principais dificultades que podemos atopar na fase de preprocesado lingüísticos.

A expansión de expresións numéricas

Unha das primeira tarefas que se debe realizar cando atopamos unha expresión numérica é identificar se se trata da expresión dun numeral cardinal ou dun numeral ordinal. Podería parecer obvio, pero non o é. Teoricamente habería que entender que cando aparece unha cifra seguida de º ou ª será un ordinal (1ª planta, 10ª edición) e que cando non é así debe ser interpretada como un número cardinal. Pero a realidade é ben distinta, e son frecuentísimos os casos en que atopamos *Planta 7* ou *12 edición*, que hai que ler respectivamente «planta sétima» e «duodécima edición».

Cómpre prever a posibilidade da expresión en números arábigos e en números romanos: *Cap. XIII* hai que lelo como «capítulo decimo terceiro».

As expresións numéricas poden ter múltiples valores:

- Poden expresar unha data: *chegou o 25-12-2013*.
- Poden expresar unha hora: *son as 7,30 da tarde*.
- Poden indicar un número de teléfono: *chámame ao 981523379*.
- Poden indicar unha cantidade dunha unidade de medida: *mide 170 metros; custou 782 euros; merquei 4 litros de aceite*.
- Pode mesmo tratarse dun número que forma parte dunha sigla: *van pechar TV3*; *fun pola A9 e volvín pola N6*.

Despois de ter identificado o valor que lles corresponde hai que asignarlles unha lectura, tendo en conta que non sempre hai unha única lectura posible. Por exemplo: *17,80 €* pódese ler:

- «dezasete euros oitenta céntimos»
- «dezasete euros con oitenta»
- «dezasete oitenta euros»
- «dezasete coma oitenta euros»

O desenvolvemento da lectura dos números de teléfono

A lectura dos número de teléfono non está normalizada. Os números de teléfono poden aparecer en grupos de dous ou tres números separados por espazos en branco, guións, puntos, ou sen agrupar. O prefixo provincial pode ter dúas ou tres cifras, e ás veces mesmo pode aparecer escrito entre parénteses. Un mesmo número de teléfono pode aparecer representado de maneiras tan distintas coma as que seguen:

- 981565283
- (981)565283
- 981.56.52.83
- 981.565.283
- 981 56 52 83
- 981 565 283

E as posibles expansións para a súa lectura tamén son múltiples, dependendo de preferencias locais, sociais e individuais. Son posibles lecturas coma as seguintes:

- «nove oito un, cinco seis, cinco dous, oito tres»;
- «novecentos oitenta e un, cinco seis cinco, dous oito tres»;
- «novecentos oitenta e un, cincuenta e seis, cincuenta e dous, oitenta e tres»;
- «novecentos oitenta e un, cincocentos sesenta e cinco, douscentos oitenta e tres»;
- etc.

Ante esta situación podemos pensar nunha posible correlación como a da seguinte táboa, pero poderíanse establecer tamén outras correlacións.

Expresión numérica	Expansión
981565283	Nove oito un, cinco seis, cinco dous, oito tres
(981)565283	Novecentos oitenta e un, cinco, seis, cinco, dous, oito, tres
981.56.52.83	Novecentos oitenta e un, cincuenta e seis, cincuenta e dous, oitenta e tres
981.565.283	Novecentos oitenta e un, cincocentos sesenta e cinco, douscentos oitenta e tres Novecentos oitenta e un, cinco seis cinco, dous oito tres
981 56 52 83	Nove oito un, cinco seis, cinco dous, oito tres

O problema da falta de estandarización no uso dos dous puntos

Os dous puntos adoitan aparecer con dous valores diferentes: tradicionalmente eran utilizados como signo de división (*240:2 = 120*), pero agora é habitual velo como indicador da separación entre as horas e os minutos (*son as 09:45*), función para a que podemos atopar tamén a utilización do apóstrofo (*son as 09'45*) e mesmo da coma (*son as 09,45*). Evidentemente, a expansión lingüística en ambos os casos ten que ser distinta: no primeiro daría lugar a unha secuencia do tipo «douscentos corenta dividido entre dous igual a cento vinte» ou «douscentos corenta entre dous igual a cento vinte»; pola contra, no segundo caso, o desenvolvemento sería algo así como «son as nove e corenta e cinco minutos» ou «son as dez menos cuarto». Para que a expansión se faga correctamente cómpre unha desambiguación previa que nos indique cal é o valor dos «:» en cada caso.

Os múltiples valores do punto

Máis complexa é aínda a expansión que debemos facer cando atopamos un punto, cando non ten un valor indicador de pausa. Vexamos algúns dos valores máis comúns que pode desempeñar:

Úsase tradicionalmente nas expresións numéricas para expresar:

a. milleiros: 3.185 debe ser desenvolvido como «tres mil cento oitenta e cinco»;
b. millóns: 3.300.000 debe ser expandido como «tres millóns trescentos/as mil». Pero, en cambio, 3.000.000 leríase «tres millóns», co que o valor do punto indicador dos milleiros sería nulo para a lectura;
c. miles de millóns: 345.260.975.700 daría lugar a «trescentos corenta e cinco mil dous centos sesenta millóns novecentos setenta e cinto mil setecentos».

Pero na norma internacional serve para separar os decimais da parte enteira: *7.48* desenvolverase como «sete con corenta e oito». Con todo, este desenvolvemento non se aplica sempre, xa que hai determinados casos que por tradición se len doutra maneira. Tal ocorre, por ex., coa lectura do valor de *pi* , a relación existente entre a lonxitude dunha circunferencia e o seu diámetro, onde 3,1416 non se le habitualmente «tres con mil catrocentos dezaseis», senón «tres catorce dezaseis».

Ás veces, na expresión do tempo, o punto separa as horas dos minutos, e unha secuencia como *chegou onte ás 12.30* podería desenvolverse como «chegou onte ás doce horas e trinta minutos» ou «chegou onte ás doce trinta» ou «chegou onte ás doce e trinta minutos», ou «chegou onte ás doce e media».

Nas datas o punto adoita utilizarse para separar a cifra ou cifras que corresponden ao día, ao mes e ao ano: *Badía Margarit morreu o 16.11.2014* hai que lelo «Badía Margarit morreu o dezaseis de novembro de dous mil catorce».

O punto é o signo, xunto co «x», que se utiliza a miúdo para indicar unha multiplicación: *13.2=26* debe lerse «trece por dous igual a vinte e seis» ou «trece multiplicado por dous igual a vinte e seis».

A expansión do guión

O guión utilízase, entre outras funcións, para:

- Indicar o intervalo entre dous números, tanto arábigos coma romanos, por ex.:

  Páxinas 14-44, caso no que a solución máis simple parece ser a substitución do guión por «a», e en consecuencia facer a expansión «páxinas catorce a corenta e catro».
  Durante os séculos V-XI, onde a solución podería ser semellante á anterior: «durante os séculos quinto a undécimo».

- Separar grupos de díxitos dos números de teléfono: 981-56-52-83, que permitiría distintos desenvolvementos dos que xa falamos, tales como: «novecentos oitenta e un, cincuenta e seis, cincuenta e dous, oitenta e tres», «novecentos oitenta e un, cinco seis, cinco dous, oito tres», «nove oito un, cinco seis, cinco dous, oito tres», «nove oito un, cincuenta e seis, cincuenta e dous, oitenta e tres».

Unha especial atención á concordancia de xénero na expansión das expresións numéricas

O galego presenta moción xenérica nos numerais cardinais *un* e *dous*, que forman o feminino en *unha* e *dúas* e, do mesmo xeito, *vinte e un / vinte e unha, sesenta e un /sesenta e unha, cento un / cento unha, mil un / mil unha*....; moción xenérica que aparece tamén na forma *centos, centas.* Por iso, cando os numerais cardinais

desempeñan unha función adxectiva, hai que prestar atención ao xénero do substantivo ao que acompañan para establecer a concordancia correspondente: *300 homes* expandirase como «trescentos homes», pero *300 mulleres* expandirase como «trescentas mulleres»; *21 euros* lerase «vinte e un euros», pero *102 pesetas* débese ler «cento dúas pesetas».

Especial atención hai que prestar aos casos en que unha expresión numérica vai seguida dunha sigla ou dunha abreviatura ou dun símbolo, coas que tamén deben concordar en xénero. Por iso secuencialmente sempre se debe expandir primeiro a abreviatura ou a sigla ou o símbolo, e despois o numeral que a acompaña para atribuírlle o xénero que lle corresponda: *2 m* debe desenvolverse «dous metros», pero a *2 p.* corresponderá «dúas páxinas» e a *2 ha* corresponderalle «dúas hectáreas».

A expansión das abreviaturas, siglas e símbolos

Para proceder á expansión das abreviaturas, das siglas e dos símbolos o sistema debe contar na fase do preprocesado lingüístico cunha lista de equivalencias, que lle permita substituír a forma abreviada pola forma ou formas desenvolvidas que se pretende obter no texto normalizado. Non todas estas expresións se comportan da mesma maneira, de aí que sexa necesario identificar o tipo de expresión para darlle a solución axeitada, sobre todo nos casos de posible ambigüidade.

Os casos de ambigüidade son moito máis frecuentes do que nunha visión superficial poida parecer:

- *col.* é a abreviatura de «colección», pero tamén o é moitas veces de «columna», e en menor medida de «colaborador» e de «color»;
- *s.* (con punto) é abreviatura de século, pero tamén o é de «seguinte»; e non debe confundirse con *s* (sen punto) que é o símbolo da unidade básica do Sistema Internacional que coñecemos como *segundo*;
- UVI é unha sigla que designa «Unidade de vixilancia intensiva», pero que tamén se utiliza para referirse ás veces a «Universidade de Vigo»;
- UCI tanto se refire a «Unidade de coidados intensivos» como a «Unión ciclista Internacional»;
- ILG refírese tanto ao «Instituto da Lingua Galega» coma ao «Instituto Lácteo Galego».

Non é este o lugar para tratar dos procedementos de desambiguación nestes casos, simplemente quero deixar apuntado que estes son fundamentalmente de dous tipos: a) os que utilizan regras ou os indicios fundados en información contextual gramatical (como a concordancia en xénero e número con adxectivos e determinantes: por ex., a presenza dun artigo feminino esixe que o elemento nuclear da unidade que representa a abreviación ou a sigla teña que ser tamén de xénero feminino); b) os que utilizan información de carácter léxico semántico extraída do conxunto do texto, que permitirá con maior probabilidade atribuír a unidade a un determinado candidato e non ao outro ou aos outros posibles: se se trata dun texto no que se fala de hospitais, é máis probable que UCI se refira a «Unidade de coidados intensivos» que a «Unión ciclista internacional».

Os símbolos e as abreviaturas deben expandirse practicamente en todos os casos. Pero non ocorre así coa lectura das siglas, que poden lerse expandidas (CCOO adoita lerse «Comisións obreiras»), pero tamén, segundo os casos, poden admitir outro tipo de lecturas:

- deletreándoas: UXT adoita lerse «u xe te»;
- léndoa coma unha palabra, cando ten unha estrutura fónica e silábica habitual na lingua: ONU lese «onu»;
- combinando deletreo e lectura silábica: CSIC lese «ce sic»

A expansión da barra

Quixera referirme, mesmo que sexa moi brevemente, á expansión da barra, que é dunha especial complexidade porque nalgún dos seus usos moi frecuentes na lingua actual dá lugar a diversos problemas de concordancia.

A barra ten en determinados usos un claro valor preposicional, en expresións do tipo: *velocidade limitada a 100 km/h , ten un soldo de* 30000 *euros/ano*, *comisión de 1,5%/ano*, que se poderían desenvolver en «velocidade limitada a cen quilómetros por hora», «ten un soldo de trinta mil euros por ano», «comisión de un con cinco por cento por ano».

Forma parte dalgunhas abreviaturas, como *r/ San Pedro,* ou *c/c*, onde a expansión será «rúa San Pedro» e «conta corrente».

Pero o problema maior da expansión da barra aparece cando esta é utilizada entre dúas palabras ou entre unha palabra e un morfema para indicar a existencia de dúas ou máis opcións posibles. Vexamos algúns casos:

• *Independentemente de que a oración estea ben/mal construída.* Neste caso a expansión non ten dificultade: «independentemente de que a oración estea ben ou mal construída»
• *Pode reclamar o/os dia/s traballado/s.* Aquí a situación xa é algo máis complexa, porque a alternancia expresada pola barra atinxe a varios elementos. Pódese optar por presentar a alternancia segregando as dúas frases nominais que hai na estrutura profunda, co que daría lugar a unha saída do tipo: «pode reclamar o día traballado ou os días traballados». Ou ben presentar a alternancia unicamente sobre o grupo de artigo + substantivo e facer concordar o adxacente unicamente sobre o grupo que vai en último lugar; neste caso a expansión sería: «pode reclamar o día ou os días traballados». Evidentemente, nos usos modernos, especialmente naqueles textos que prestan unha especial atención á non discriminación de xénero na linguaxe, aparecen situacións que moi dificilmente poden recibir unha solución satisfactoria.

A existencia de varias alternativas de lectura

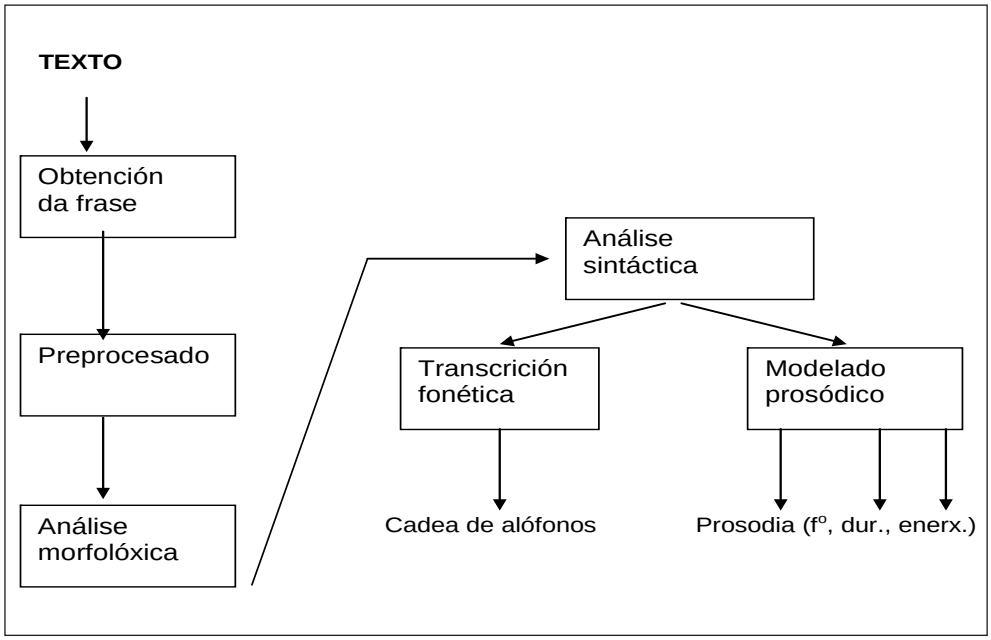
Vimos que existen casos en que son perfectamente posibles distintas lecturas correctas nas expansións que se realizan no preprocesado lingüístico: *eran as 12:30* pode desenvolverse como «eran as doce trinta» ou «eran as doce horas e trinta minutos» ou «eran as doce e media».

Cando hai varias alternativas de lectura, pódese optar por darlle prioridade a unha delas e atribuír sempre a mesma saída a un elemento que hai que desenvolver; pódese optar por aceptar todas as posibles expansións e aplicalas aleatoriamente, co que se logra unha maior variedade estilística; ou pódese mesmo atribuír unha solución ou outra en función de determinadas regras ou condicións (*se se cumpre a condición A, a saída será «x»; se se cumpre a condición B, a saída será «y»; se se cumpre a condición C, a saída será «z»; se non se cumpre ningunha das condicións anteriores, a saída será «v»*).

Cando debe facerse o preprocesado lingüístico

É evidente que o preprocesado lingüístico debe ser realizado nas fases iniciais do proceso que desembocará na transformación do texto escrito en texto oral. Pero a pregunta clave é: debe realizarse o preprocesado lingüístico antes da segmentación do texto en enunciados (o que se coñece como fase de *obtención da frase*) ou despois?

Frecuentemente o tratamento lingüístico dun sistema de síntese de voz preséntase de acordo co seguinte diagrama:



Nesta secuenciación o primeiro paso que se dá é a segmentación do texto en enunciados (o que no diagrama se representa como *obtención de frase*), o que implica que a normalización do texto levada a cabo na fase do preprocesado lingüístico faise de maneira independente sobre cada un dos enunciados, e non previamente sobre a totalidade do texto que se vai tratar.

Pode parecer indiferente que este proceso se faga dunha ou outra maneira, pero non o é. A aplicación de regras de concordancia de xénero e número na maioría dos casos pódese realizar acudindo unicamente ao contexto existente nun enunciado; algunhas regras estatísticas poden ser aplicadas tamén dentro deste marco, outras sono menos eficientemente. Pero o problema xorde de maneira moi importante cando se pretende botar man para a desambiguación de determinados elementos de información léxico-semántica existentes no texto, pero que raramente se atopan no ámbito tan estreito da cadea lingüística comprendida entre dous puntos.

Conclusión

Esta achega non ten como finalidade afondar nos procedementos para resolver todas as situacións problemáticas que se presentan na fase de preprocesado lingüístico nos sistemas de síntese de voz. Simplemente pretende chamar a atención sobre o feito de que este módulo inicial de regularización ou de normalización do texto non está exento de problemas, que deben ser estudados con exhaustividade, porque a resolución acertada ou non destes vai condicionar dunha maneira importante o resultado final do proceso.