Gómez Guinovart, Xavier (2000): "Lingüística computacional".
En Ramallo, Fernando, Gabriel Rei-Doval e Xoán Paulo Rodríguez (eds.),
Manual de ciencias da linguaxe, cap. 6 (pp. 221-268). Vigo: Xerais.

LINGÜÍSTICA COMPUTACIONAL

Xavier Gómez Guinovart
Universidade de Vigo

1. ÁMBITO DA LINGÜÍSTICA COMPUTACIONAL

1.1. Liñas de investigación

A *lingüística computacional* constitúe un eido científico interdisciplinario vinculado á lingüística e á informática, e encamiñado a incorporar nos ordenadores a habilidade no manexo da linguaxe natural humana e a facilita-lo tratamento informatizado das linguas e do seu estudio. A pesar de constituír unha disciplina relativamente recente, unha aproximación tentativa á delimitación do seu campo de estudio esixe o recoñecemento dun mínimo de tres liñas principais de investigación e desenvolvemento: a lingüística computacional teórica, a lingüística computacional aplicada e a informática aplicada á lingüística.

A primeira vertente da disciplina, a *lingüística computacional teórica*, está constituída pola que se considera a lingüística computacional por antonomasia. Dentro desta liña de investigación cómpre distinguirmos alomenos tres obxectivos complementarios: a elaboración de modelos lingüísticos en termos formais e implementables, a aplicación destes modelos a calquera nivel de descrición lingüística e a comprobación automatizada da congruencia dunha teoría lingüística e das súas prediccióǹs. A confección de modelos computacionais da linguaxe mediante formalismos lingüísticos axeitados e a utilización destes modelos na descrición lingüística facilitan a detección de erros e incoherencias na descrición dos fenómenos lingüísticos e ofrecen un medio práctico e efectivo para observa-la interacción de tódolos compoñentes do modelo lingüístico teórico postulado. Na segunda sección do capítulo, trataremos de presentar algúns dos aspectos máis cultivados desta faceta do estudio.

En segundo lugar, a orientación máis tecnolóxica da lingüística computacional, a *lingüística computacional aplicada*, diríxese ó deseño e elaboración de sistemas informáticos capaces de comprender, producir e traducir enunciados orais e escritos en linguas naturais; e concrétase no desenvolvemento de aplicacións lingüísticas da informática nas que se poden distinguir catro grandes categorías: os sistemas de comprensión e xeración de enunciados (como os programas de consulta en linguaxe natural a bases de datos e os sistemas automáticos de diálogo por liña telefónica), as aplicacións das tecnoloxías da fala (como os programas de dictado e os sistemas de

conversión de texto a voz), as ferramentas de procesamento documental para a elaboración, xestión e revisión de documentos textuais (como os programas de verificación da corrección lingüística dos textos, os programas de xeración automática de resumos, os sistemas de extracción de información, os sistemas de recuperación da información textual e os programas de catalogación documental automatizada) e, en derradeiro lugar, as ferramentas de procesamento plurilingüe, na súa dobre vertente de aplicacións didácticas para o ensino de linguas (como os métodos de aprendizaxe de idiomas asistida por ordenador e os programas de creación de exercicios de lingua) e de ferramentas de axuda á traducción (como os programas de traducción automática, as bases de datos terminolóxicos e as utilidades de memoria de traducción). Dependendo da faciana desta actividade que se pretenda salientar, este campo de traballo recibe as denominacións de *procesamento da linguaxe natural*, *tecnoloxías da lingua* ou *enxeñería lingüística*. Na alínea terceira do capítulo, examinaremos algunhas das aplicacións máis desenvolvidas nesta dirección, como son o recoñecemento e a síntese da fala, a extracción da información textual, a verificación da corrección lingüística dos textos e a traducción automática.

Por último, o campo de traballo caracterizado pola aplicación dos ordenadores á investigación lingüística, é dicir, ó estudio científico da linguaxe e das linguas, adoita recibi-lo nome de *informática aplicada á lingüística* ou, con maior concisión, de *lingüística informática*. O termo pode aplicarse en sentido amplo a tódalas subdisciplinas da lingüística que empregan ferramentas informáticas, aínda que en xeral se reserva o seu emprego para as áreas de investigación onde estas ferramentas teñen unha maior incidencia, como a lingüística de corpus ou a lingüística histórica computacional, dúas áreas específicas de estudio que se revisarán na cuarta alínea deste capítulo.

1.2. Dimensións interdisciplinaria e social

Desde o punto de vista da súa vinculación coa informática, e tamén por motivos históricos, a lingüística computacional está considerada unha subdisciplina da *intelixencia artificial*, unha especialidade da informática que se ocupa da comprensión da intelixencia e do deseño de máquinas intelixentes, é dicir, de máquinas e programas que presentan características asociadas co entendemento humano, como o raciocinio, a comprensión da linguaxe falada e escrita, a aprendizaxe ou a toma de decisións.

Así mesmo, desde o punto de vista da súa ligazón coa lingüística, a lingüística computacional debe considerarse tamén unha subdisciplina da *lingüística teórica*, xa que un dos seus obxectivos é a elaboración de modelos formais e implementables da linguaxe humana. Neste sentido, a lingüística computacional está estreitamente relacionada coa *psicolingüística* e coa *lingüística cognitiva*, polo seu interese compartido na descrición e modelado da actividade mental implicada no procesamento lingüístico.

Finalmente, como disciplina lingüística experimental, a lingüística computacional constitúe a área de traballo da *lingüística aplicada* especificamente interesada en aplica-los resultados e métodos da investigación lingüística á elaboración de productos comerciais e de investigación no marco das industrias da lingua. O amplo abano de aplicacións lingüísticas da informática enlaza a lingüística computacional coas diferentes disciplinas lingüísticas e non lingüísticas relacionadas con cada unha das aplicacións, como a *enxeñería de telecomunicacións* (en relación coas aplicacións das

tecnoloxías da fala), as *ciencias da documentación* (cos sistemas de xestión documental), a *traductoloxía* (coas ferramentas de axuda á traducción), a *didáctica das linguas* (co ensino de linguas asistido por ordenador), a *análise do discurso* (cos sistemas de diálogo) ou a *lexicografía* (cos diccionarios electrónicos).

Xunto a esta dimensión interdisciplinaria da disciplina, cómpre tamén salientármo-la súa dimensión social, nun contorno configurado por unha sociedade da información cada vez máis global e máis entretecida polas telecomunicacións e polos intereses culturais e comerciais que estas sustentan. As aplicacións actuais da lingüística informática ás telecomunicacións manifestan unha clara tendencia a permitiren o uso da lingua propia de cadaquén para acceder a tódalas posibilidades de información, comunicación e consumo postas á disposición das persoas habitantes do denominado "primeiro mundo" polas novas tecnoloxías. Pola súa grande incidencia social, a presencia dunha lingua neste ámbito ha ser determinante para acadar ou para conserva-lo estado de lingua normalizada (Comisión Europea 1998, 14-15).

2. MODELOS E FORMALISMOS LINGÜÍSTICOS

A implementación informática de teorías lingüísticas e a elaboración de modelos computacionais da linguaxe constitúen intereses centrais da investigación en lingüística computacional. Preséntanse a continuación algunhas das orientacións máis salientables destas dúas liñas de traballo complementarias, limitándonos ós niveis lingüísticos do léxico e da sintaxe, e ós modelos simbólicos e formalismos de unificación. Outras liñas destacadas de traballo en lingüística computacional na actualidade son a fonoloxía computacional (Bird 1995), as redes léxico-semánticas (Wanner 1996, Alonge et al. 1998), a semántica computacional (Rosner e Johnson 1992) e os modelos lingüísticos probabilísticos (Charniak 1993).

2.1. Modelos lingüísticos computacionais

Un dos obxectivos fundamentais da lingüística computacional é o desenvolvemento de teorías lingüísticas formais e implementables informaticamente. Estas teorías constitúen así modelos computacionais do funcionamento da linguaxe, razón pola que se denominan *modelos lingüísticos* (Shieber 1988). Dentro desta liña de investigación, deben mencionarse modelos lingüísticos computacionais como a gramática léxica funcional ou LFG (Bresnan 1999), a gramática sintagmática xeneralizada ou GPSG (Gazdar, Klein, Pullum e Sag 1985, Bennett 1995), a gramática sintagmática dirixida polo núcleo ou HPSG (Pollard e Sag 1994, Borsley 1996) e a gramática categorial (Solias 1996), modelos agrupados xenericamente baixo a denominación de *gramáticas de unificación* (Shieber 1986, Ruiz 1996, Balari 1999) polo recurso a este procedemento matemático nas súas descricións lingüísticas.

Na maioría destes modelos, os obxectos lingüísticos están representados en forma de obxectos matemáticos denominados *estructuras de trazos* (ET) e os fenómenos lingüísticos descríbense formulando ecuacións con estas ET. A unificación é a operación matemática que permite resolver estes sistemas de ecuacións, combinando as informacións lingüísticas codificadas nas ET asociadas. Cada ET está formada por unha matriz de trazos, e cada trazo consiste nunha parella *[Atributo=Valor]* que especifica un parámetro lingüístico e o valor que adopta ese parámetro, por exemplo:

[número=singular] ou *[persoa=terceira]*. Así, un xeito de representa-la información lingüística asociada co pronome *ela* sería mediante a ET da figura 1.

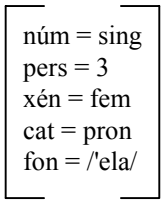


Figura 1. Estructura de trazos para o pronome *ela*.

O valor dun trazo pode ser outro trazo, converténdose daquela nun trazo complexo. Por exemplo, a oración *"o neno come a sopa"* pódese representar coa ET da figura 2, onde os trazos para o suxeito e para o complemento directo son trazos complexos.

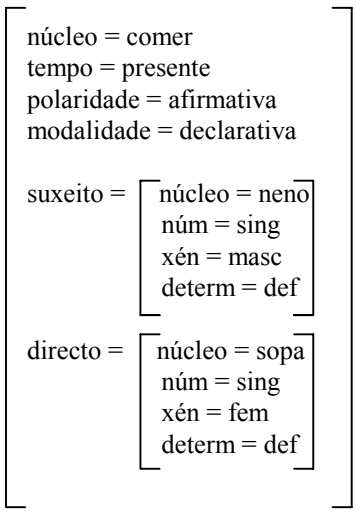


Figura 2. Estructura de trazos oracional con trazos complexos.

A unificación é unha operación que combina dúas ET e produce como resultado outra ET que posúe toda a información contida nas dúas ET orixinais, sempre que a información non sexa contradictoria. Se a información contida nunha das ET resulta contradictoria coa contida na outra, non pode realizarse a unificación. Por exemplo, na figura 3, a ET1 e a ET2 unifican en ET4, pero a ET3 non pode unificar con ET1 nin con ET2.

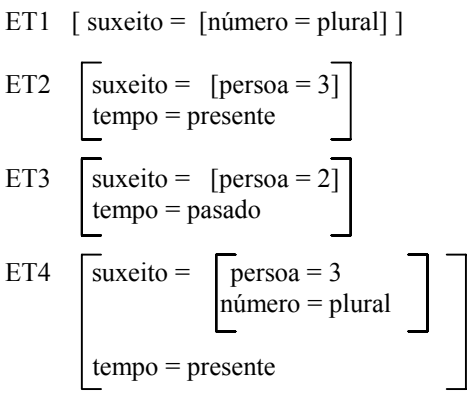


Figura 3. Exemplo de unificación de estructuras de trazos.

Por outra banda, na formalización de modelos lingüísticos, o procedemento máis espallado de descrición das estructuras de constituíntes admitidas nunha lingua son as regras sintagmáticas (tamén denominadas regras de estructura de constituíntes ou regras de reescritura). Estas regras adoptan a forma A→B, onde A representa unha categoría sintáctica e B son os constituíntes inmediatos de A. Por exemplo, unha regra que expresaría a estructura sintagmática de moitas oracións do galego podería ser O→SN^SV ("unha oración está formada por un sintagma nominal seguido dun sintagma verbal"). Basicamente, unha gramática de estructura sintagmática é un conxunto de regras sintagmáticas que describen as estructuras sintácticas aceptables nunha lingua. O exemplo da figura 4 podería constituír un fragmento dunha gramática sintagmática do galego que describiría, entre outras, a estructura de constituíntes representada mediante un diagrama arbóreo na figura 5.

O → SN SV
SN → Det N
SV → V SN

Det → *o*
Det → *os*
N → *ordenador*
N → *ordenadores*
N → *texto*
N → *textos*
V → *procesa*
V → *procesan*

Figura 4. Fragmento de gramática de estructura sintagmática.

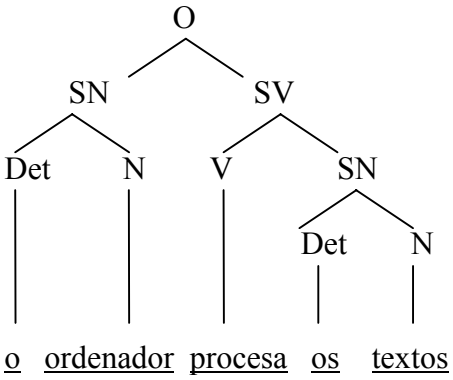


Figura 5. Estructura de constituíntes descrita pola gramática.

As regras sintagmáticas aumentadas permiten caracteriza-los constituíntes e establecer condicións de igualdade entre as súas propiedades. Por exemplo, pódese aumentar con condicións o fragmento anterior de gramática sintagmática do galego, para incorpora-la obrigatoriedade da concordancia de número e persoa entre o verbo e o seu suxeito, modificando a primeira regra como se mostra na figura 6.

$$O \rightarrow SN\ SV$$
$$\langle SN\ núm \rangle = \langle SV\ núm \rangle$$
$$\langle SN\ per \rangle = \langle SV\ per \rangle$$

Figura 6. Regra sintagmática aumentada.

A interpretación desta regra indicaría que unha oración está formada por un SN seguido dun SV, e que o número e a persoa do SN e do SV coinciden. Substituíndo os símbolos categoriais indivisibles por estructuras de trazos, como na figura 7, pódense redefinir estas ecuacións en forma de igualdade de variables.

$$[cat = O] \rightarrow \begin{bmatrix} cat = SN \\ núm = \alpha \\ per = \beta \end{bmatrix}, \begin{bmatrix} cat = SV \\ núm = \alpha \\ per = \beta \end{bmatrix}$$

Figura 7. Estructuras de trazos e regras sintagmáticas aumentadas.

Finalmente, asignándolle ó SN a función de suxeito e ó SV a de predicado, e empregando estas dúas funcións como atributos de trazos complexos, poderíase reduci-la regra a unha simple estructura de trazos (Figura 8).

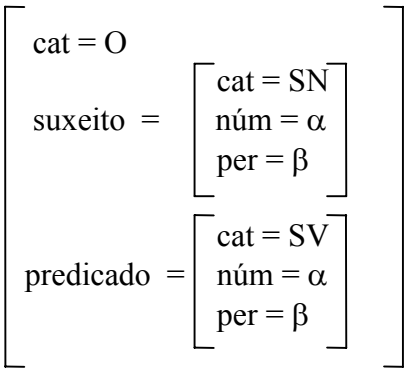


Figura 8. Estructura de trazos con variables.

Esta perspectiva estática da manipulación dos obxectos lingüísticos permite simplifica-lo compoñente sintáctico e desprazar cara ó léxico a información lingüística tradicionalmente tratada na gramática, o que redunda nunha maior complexidade da organización e contido do compoñente léxico. O resultado práctico deste desprazamento é que nos modelos de orientación lexicista as regras sintagmáticas son inexistentes (como na gramática categorial) ou de natureza moi xeral (como na HPSG), o que implica habitualmente a adopción dalgunha das versións da teoría da X-barra, onde o núcleo do constituínte, caracterizado xa no léxico, proporciona practicamente toda a información sobre as propiedades sintácticas e semánticas dos seus complementos. Unha simplificación adicional da gramática consiste en distinguir entre as regras que representan a orde secuencial dos constituíntes (regras de precedencia lineal) e as que representan as súas relacións de dependencia estructural ou xerárquica (regras de dominio inmediato), distinción practicada e difundida pola GPSG e a HPSG.

O compoñente léxico destes modelos organízase como unha rede de nós xerarquizada con herdanza múltiple, onde cada nó está formado por unha estructura de trazos correspondente a un lexema ou a unha clase de lexemas. Os nós para as clases de lexemas conteñen tódalas propiedades lingüísticas compartidas pola clase. Por exemplo, a clase "verbos regulares da primeira conxugación" do léxico galego podería conte-la información morfolóxica necesaria para flexiona-los verbos desta clase; a clase "verbos transitivos", a especificación do contexto sintáctico característico dos verbos desta categoría; e a clase "verbos", a adscrición categorial compartida por tódolos verbos do idioma. Os nós das redes léxicas poden herdar unha parte das súas propiedades dos nós da xerarquía dos que dependen. Por exemplo, o nó da clase "verbos transitivos" pode herda-la súa categoría sintáctica do nó da clase "verbos", e o nó do lexema *tomar* pode herda-las súas propiedades categoriais e de subcategorización do nó da clase "verbos transitivos". Ademais, a herdanza pode vir de diferentes nós, sempre que as propiedades herdadas non sexan contradictorias. Así, o nó do lexema *tomar* podería herda-las súas características sintácticas do nó da clase "verbos transitivos" e as súas características morfolóxicas da clase "verbos regulares da primeira conxugación" (Figura 9). Deste xeito, conséguese eliminar do léxico a información lingüística redundante, xa que non é necesario repeti-la descrición da subcategorización transitiva en tódalas entradas dos verbos transitivos, nin o comportamento flexivo da primeira conxugación en tódalas entradas dos verbos deste paradigma morfolóxico.

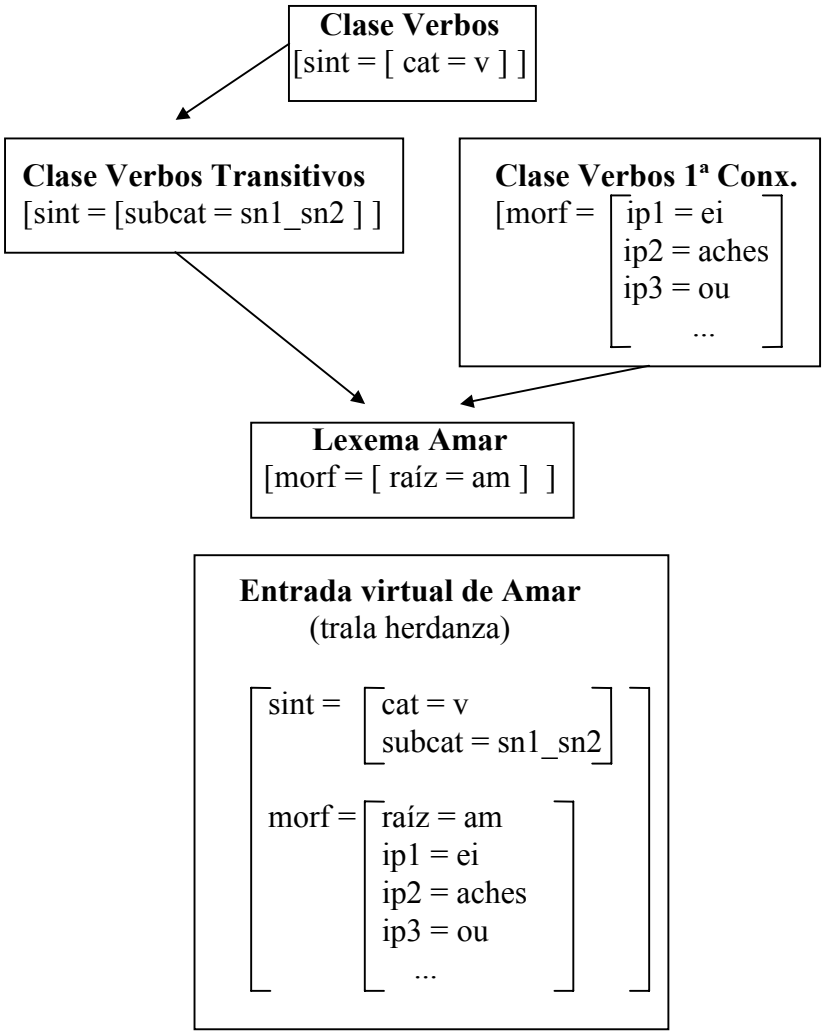


Figura 9. Representación da información léxica.

2.2. Formalismos lingüísticos

Os *formalismos lingüísticos* (ou *sistemas de programación lingüística*) son linguaxes artificiais deseñadas para representa-la información lingüística. Algúns formalismos lingüísticos —como DCG (Pereira e Warren 1980), FUG (Kay 1982), PATR (Shieber 1986), DATR (Evans e Gazdar 1996), a morfoloxía de dous niveis (Koskenniemi 1983) ou ALE (Carpenter e Penn 1997)— tamén son entendidos (ou, con maior exactitude, interpretados) polos ordenadores, polo que son especialmente adecuados para a implementación informática e a comprobación automática das teorías lingüísticas. Nestas actividades, como complemento ou substituto dos formalismos lingüísticos, empréganse tamén linguaxes de programación de propósito xeral e, en particular, a linguaxe de programación Prolog (Gazdar e Mellish 1989). Preséntanse seguidamente dúas aplicacións simples, ilustrativas dos métodos de programación

lingüística nos niveis de análise sintáctica e morfolóxica, mediante os formalismos PATR e DATR.

PATR está deseñado para escribir gramáticas de estructura sintagmática aumentadas con estructuras de trazos sobre as que opera a unificación. PC-PATR[1] é un programa que permite implementar informaticamente gramáticas escritas neste formalismo. Por exemplo, para converter a PC-PATR o fragmento anterior de gramática de estructura sintagmática do galego (co engadido dalgunhas comprobacións da concordancia nominal e verbal), primeiro cómpre crear un ficheiro que conteña a gramática, ficheiro que debe te-la extensión *.grm*, como en *patr01.grm* (Figura 10).

```
Rule O -> SN SV
      <SN num> = <SV num>

Rule SV -> V SN
      <SV num> = <V num>

Rule SN -> Det N
      <SN num> = <N num>
      <Det xen> = <N xen>
      <Det num> = <N num>
```

Figura 10. Gramática *patr01.grm* en PC-PATR.

En PC-PATR, as regras deben ir precedidas da palabra inglesa *Rule*. A primeira regra do exemplo define unha estructura de constituíntes aumentada con trazos, na que opera a unificación sobre os trazos de número nominal e verbal (Figura 11). Se estes trazos fosen contradictorios, non se podería aplica-la unificación e, por tanto, non se podería realiza-la análise da secuencia.

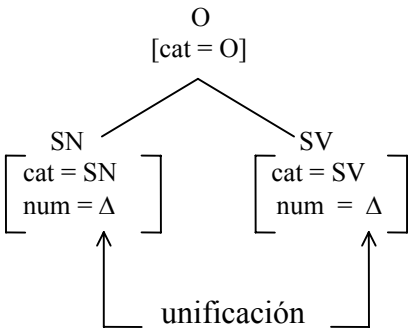


Figura 11. Unificación do número nominal e verbal.

Na estructura sintáctica definida pola segunda regra, a unificación aplícase ós trazos de número verbal, de maneira que o número do verbo sexa tamén o número do seu sintagma verbal (Figura 12). Mediante esta elevación do valor de número de V a

[1] ftp://ftp.sil.org/software/dos/pcp099b5.zip.

SV, obtense un SV co número de acordo co seu núcleo, o que permite a comprobación da concordancia co suxeito expresada na primeira regra.

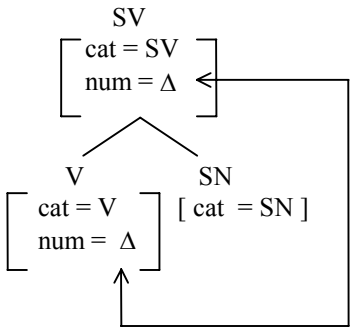


Figura 12. Elevación de número por unificación.

Finalmente, na terceira regra, constrúese a estructura de constituíntes do SN, elévase o valor de número do núcleo nominal ó SN resultante, e compróbase a concordancia de xénero e número entre o determinante e o nome (Figura 13).

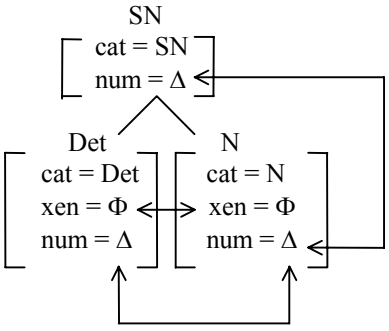


Figura 13. Concordancia e elevación de número.

Unha vez elaborada a gramática e almacenada en *patr01.grm*, deben definirse as regras que introducen as pezas léxicas terminais nun ficheiro coa extensión *.lex*, por exemplo, *patr01.lex* (Figura 14).

```
\w o                    \w texto
\c Det                  \c N
\f <xen> = m            \f <xen> = m
   <num> = s               <num> = s

\w os                   \w textos
\c Det                  \c N
\f <xen> = m            \f <xen> = m
   <num> = p               <num> = p
```

```
\w ordenador             \w procesa
\c N                     \c V
\f <xen> = m             \f <num> = s
   <num> = s

\w ordenadores           \w procesan
\c N                     \c V
\f <xen> = m             \f <num> = p
   <num> = p
```

Figura 14. Léxico *patr01.lex* en PC-PATR.

A partir destas definicións créanse unhas estructuras de trazos dispostas para a súa inserción nunha estructura sintagmática. Estas ET léxicas conterán a información declarada en PC-PATR tralo código *\w* (por *word* "palabra") convertida no valor de *lex* (por lexema), a declarada tralo código *\c* convertida no valor de *cat*, e mais tódolos outros trazos especificados tralo código *\f* (por *features* "trazos"), como se mostra na ET da figura 15 para a palabra *ordenadores*.

$$\begin{bmatrix} \text{lex} = \text{ordenadores} \\ \text{cat} = \text{N} \\ \text{xen} = \text{m} \\ \text{num} = \text{p} \end{bmatrix}$$

Figura 15. Estructura de trazos léxica en PC-PATR.

Dada a gramática de *patr01.grm* e o léxico de *patr01.lex*, unha interacción típica co sistema pode se-la que se recolle na figura 16, onde PC-PATR realiza a análise sintáctica automática da secuencia *"o ordenador procesa os textos"* e, como resultado, constrúe a súa estructura de constituíntes e indica as ET asociadas con cada nó (numerado, para maior claridade) da árbore sintáctica.

```
PC-PATR>load grammar patr01
Loading grammar from patr01.grm
PC-PATR>load lexicon patr01
Loading lexicon from patr01.lex
    8 lexicon entries loaded from patr01.lex
PC-PATR>parse
Sentence: o ordenador procesa os textos
1:
               O_1
       _________|__________
     SN_2                SV_5
   ____|_____          _____|_____
 Det_3     N_4      V_6      SN_7
  o    ordenador procesa    ___|____
                          Det_8   N_9
                           os  textos
```

```
O_1:
[ cat:   O ]
SN_2:
[ cat:   SN
  num:   s ]
Det_3:
[ cat:   Det
  xen:   m
  lex:   o
  num:   s ]
N_4:
[ cat:   N
  xen:   m
  lex:   ordenador
  num:   s ]
SV_5:
[ cat:   SV
  num:   s ]
V_6:
[ cat:   V
  lex:   procesa
  num:   s ]
SN_7:
[ cat:   SN
  num:   p ]
Det_8:
[ cat:   Det
  xen:   m
  lex:   os
  num:   p ]
N_9:
[ cat:   N
  xen:   m
  lex:   textos
  num:   p ]
1 parse found
```

Figura 16. Análise sintáctica automática con PC-PATR.

Por outra banda, cara á programación lingüística de léxicos computacionais, a linguaxe formal DATR permite implementar informaticamente redes de estructuras de trazos léxicas (de lexemas e clases de lexemas) xerarquizadas e con herdanza de propiedades múltiple (é dicir, herdanza que pode vir de diferentes nós, como se explica na alínea anterior). Así, a figura 17 recolle o exemplo de rede léxica do galego ilustrado na figura 9, convertido a DATR e almacenado nun ficheiro (aquí denominado *herdanza.dtr*) para o seu posterior procesamento.

```
Verbos:
 <sint cat> == v.
Verbos_Transitivos:
 <> == Verbos
 <sint subcat> == sn1_sn2.
Verbos_1Conx:
 <morf ip1> == ei
 <morf ip2> == aches
 <morf ip3> == ou.
```

```
Tomar:
 <morf raiz> == tom
 <sint> == Verbos_Transitivos
 <morf> == Verbos_1Conx.
#show
<sint cat> <sint subcat> <morf raiz> <morf ip1> <morf ip2>
<morf ip3>.
#hide
Verbos Verbos_Transitivos Verbos_1Conx.
```

Figura 17. Ficheiro de léxico *herdanza.dtr* en DATR.

De acordo coas convencións de DATR, en *herdanza.dtr* os nomes dos nós van seguidos de dous puntos (:), os trazos da súa ET aparecen entre os dous puntos e un punto final (.), os atributos (ou cadeas de atributos) das estructuras de trazos van entre corchetes triangulares (<>) e os seus valores indícanse con dous signos de igual (==). A presencia do nome dun nó como valor dun atributo sinala a orixe da herdanza; por exemplo, o trazo *<morf>==Verbos_1Conx* do nó *Tomar* indica que este nó herda do primeiro as propiedades do atributo *<morf>*. Os atributos baldeiros (<>) empréganse para establece-la herdanza de tódolos trazos incluídos no nó especificado como valor; así, a liña *<>==Verbos* do nó *Verbos_Transitivos* significa que este nó herda tódolos trazos da ET do nó *Verbos* (neste caso, *<sint cat>==v*). Finalmente, tralo código *#show* aparecen os nomes dos atributos que se pretenden visualizar como resultado do tratamento do ficheiro, e tralo código *#hide* xusto o contrario, pero en referencia ós nomes dos nós. Con esta definición do léxico e esta configuración, e utilizando un programa como Q-DATR[2] para o seu procesamento informático, o ordenador pode realiza-la avaliación automática da herdanza de propiedades especificada no léxico, e presentar seguidamente os trazos dos nós da rede solicitados (Figura 18). Está dispoñible unha implementación completa en DATR da morfoloxía flexiva verbal do galego, distribuída en Internet polo Seminario de Lingüística Informática da Universidade de Vigo[3].

```
>> cp(c:\datr\herdanza.dtr)
compiling c:\datr\herdanza.dtr
6 sentences compiled
>> datr_theorem
Tomar:
          <sint cat> = v
          <sint subcat> = sn1_sn2
          <morf raiz> = tom
          <morf ip1> = ei
          <morf ip2> = aches
          <morf ip3> = ou.
```

Figura 18. Resolución da herdanza con Q-DATR.

[2] ftp://ftp.cogs.sussex.ac.uk /pub/nlp/DATR/qdatr200.exe.
[3] http://www.uvigo.es/webs/sli/.

3. APLICACIÓNS DA LINGÜÍSTICA COMPUTACIONAL

3.1. Comprensión e xeración de linguaxe natural

Un dos obxectivos centrais da lingüística computacional aplicada é permiti-lo uso oral da lingua materna como medio de comunicación entre os ordenadores e as persoas, coa finalidade de que as persoas poidan acceder a tódalas facilidades ofrecidas polos ordenadores mediante ordes vocais expresadas espontaneamente co vocabulario e a sintaxe da súa propia lingua e, ó mesmo tempo, que os ordenadores presenten os resultados das súas aplicacións nesa mesma lingua de maneira natural e inmediatamente comprensible para as persoas. Os sistemas de comprensión de linguaxe natural son os programas informáticos que se encargan de deduci-lo significado dos enunciados lingüísticos de entrada que procesan, mentres que os sistemas de xeración da linguaxe natural son os responsables de presenta-los resultados das aplicacións informáticas en forma de enunciados lingüísticos (Allen 1995, Reiter e Dale 1997). A combinación das técnicas de comprensión e xeración permite o establecemento dunha interacción lingüística entre persoa e ordenador en situacións comunicativas ben delimitadas, como as que se dan nos programas de consulta en linguaxe natural a bases de datos ou nos sistemas automáticos de diálogo por liña telefónica (Figura 19).

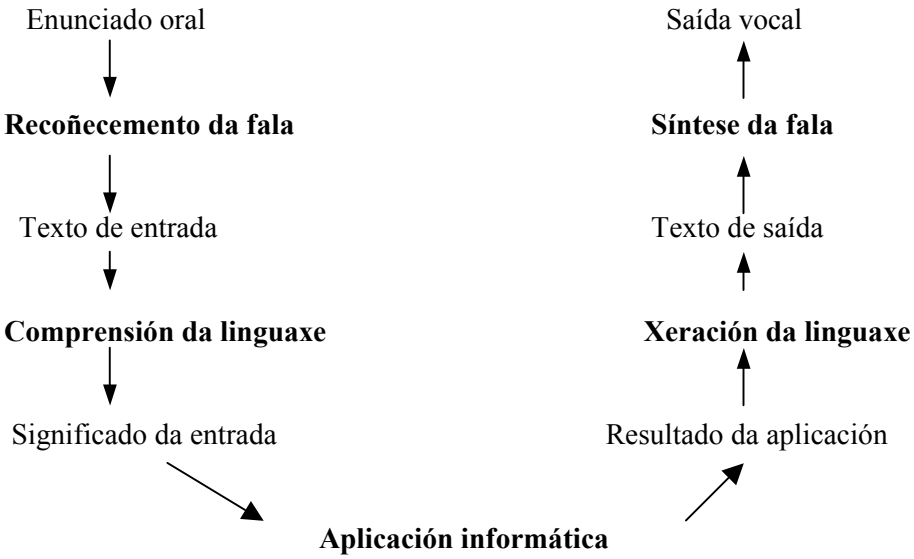


Figura 19. Interacción lingüística oral persona-ordenador.

A xeración e a comprensión da linguaxe natural son tarefas complexas que implican a aplicación conxunta de moi diversas técnicas de análise e producción lingüística automática. Deixando á marxe o procesamento do nivel fónico, que examinaremos no seguinte apartado (dedicado especificamente ó recoñecemento e síntese da fala), o procesamento da comprensión lingüística realízase habitualmente en catro etapas sucesivas, correspondentes á análise morfolóxica (etiquetado categorial e lematización), a análise sintáctica, a análise semántica oracional e a análise pragmática e discursiva; mentres que a tarefa da xeración de linguaxe percorrería o camiño inverso,

seguindo as etapas de planificación semántica, planificación sintáctica, selección léxica e xeración morfolóxica (Figura 20).

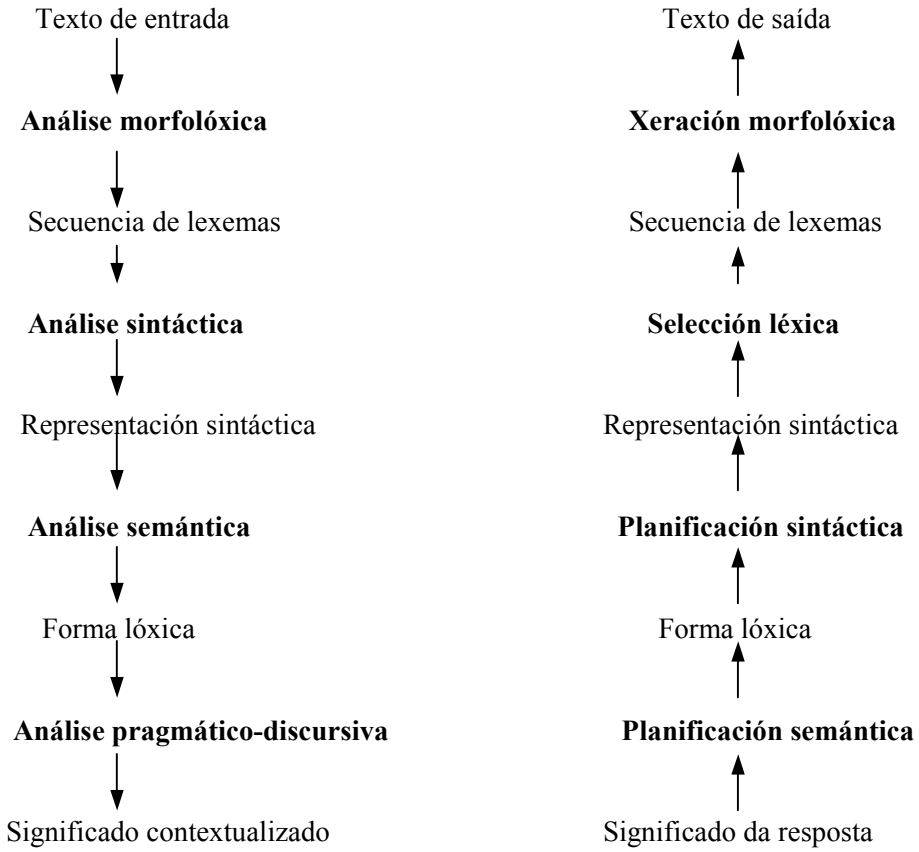


Figura 20. Compoñentes da comprensión e xeración da linguaxe.

A utilidade operativa das aplicacións de comprensión e xeración da linguaxe está actualmente circunscrita a ámbitos de interacción moi ben delimitados desde un punto de vista temático e lingüístico, como pode se-la consulta telefónica da base de datos dos horarios dos voos dunha compañía aérea concreta mediante diálogos dirixidos polo ordenador. Cómpre seguir traballando para que o procesamento da comprensión da linguaxe acade uns niveis elevados de cobertura e precisión capaces de manexar axeitadamente os enunciados que aparecen nas interaccións lingüísticas espontáneas. Un nivel de cobertura alto implica que o programa de comprensión non deixe case ningún enunciado sen analizar, mentres que un grao alto de precisión supón que a maioría dos enunciados analizados reciben unha análise correcta. Unha cobertura apropiada evitaría que a persoa usuaria do sistema tivese que repetir de distintas maneiras un enunciado por indicación do programa, mentres que unha boa precisión evitaría os erros de interpretación por parte do sistema.

No eido da xeración da linguaxe, a investigación está centrada sobre todo na xeración de linguaxe escrita e nos contornos comunicativos de interactividade baixa. Para poder dispoñer nun futuro de sistemas con interacción lingüística oral espontánea persoa-ordenador, será preciso orienta-los esforzos de investigación cara á xeración de linguaxe oral en diálogos interactivos, nos que o programa de xeración teña en conta o

contido dos enunciados previos e adapte a súa producción lingüística ás intervencións da persoa interlocutora.

3.2. Tecnoloxías da fala

As tecnoloxías da fala ocúpanse do procesamento dos aspectos fónicos da linguaxe, co obxectivo de permitiren a comunicación oral entre as persoas e os ordenadores. Segundo a dirección da comunicación considerada, o tratamento informático da fala afronta dúas tarefas ben diferenciadas: o procesamento da percepción acústica ou *recoñecemento da fala*, e o procesamento da producción fonética ou *síntese da fala* (Dutoit 1997, Llamas e Cardeñoso 1997).

O recoñecemento da fala consiste en converter un enunciado oral nunha cadea de símbolos, por exemplo, nun texto escrito. A popularización das tecnoloxías de recoñecemento débese ós sistemas de dictado para procesamento de texto en ordenadores persoais. Estes programas de dictado, comercializados por empresas como IBM e Dragon Systems, ofrecen versións para fala fragmentada, nas que se debe facer unha pausa entre as palabras, e versións para fala continua, que permiten dictar texto sen necesidade de facer pausas entre as palabras.

Unha das características máis desexables nun sistema de recoñecemento é a resistencia ó ruído do ambiente, coa finalidade de podelo utilizar en contornos ruidosos (como nunha fábrica, para controlar vocalmente o brazo dun robot) ou a través do teléfono (por exemplo, para dictarlle a unha central telefónica automatizada o número do abonado ó que se pretende chamar). Polo de agora, malia o interese evidente que esta cuestión suscita entre os provedores de servicios de telecomunicacións, aínda non hai unha solución definitiva para a baixa fiabilidade do recoñecemento en contornos ruidosos (Tapias 1999).

Outra das dificultades do recoñecemento irrestricto da fala continua consiste en recoñece-la fala con independencia da persoa. Un recoñecedor é irrestricto cando recoñece o vocabulario xeral dunha lingua. Isto é imprescindible nun sistema de dictado, aínda que outras aplicacións, como as centrais telefónicas automatizadas, poden limitarse a recoñecer unhas poucas palabras. Así mesmo, unha central automatizada dun sistema público de consulta telefónica debe ser capaz de recoñece-la fala de calquera que chame, mentres que un sistema de dictado pode especializarse nas características da fala dunha persoa concreta. Como o recoñecemento irrestricto de fala continua con independencia da persoa aínda non atinxiu un grao de fiabilidade aceptable para a súa comercialización, os sistemas de dictado para fala continua requiren unha fase de adestramento, consistente nunha media hora de lectura individual dun texto preparado, na que o sistema adquire os datos necesarios sobre as características acústicas da voz e sobre a pronunciación particular dos sons da lingua.

A síntese da fala fai o camiño inverso ó do recoñecemento: a conversión de cadeas de símbolos en enunciados orais. Por exemplo, nun sistema de síntese para invidentes que vocalice o texto da pantalla dun ordenador persoal, a cadea de símbolos convertida polo sintetizador son as letras agrupadas en palabras e acompañadas por signos de puntuación.

No estado actual do desenvolvemento tecnolóxico, a intelixibilidade da emisión sonora, premisa básica da voz sintetizada, é un problema xa resolto de maneira case que definitiva. Sen embargo, aínda cómpre soluciona-la cuestión da naturalidade da pronuncia, é dicir, conseguir que a fala xerada polo ordenador non soe a voz de robot. A

clave para acadar este obxectivo podería se-la curva de entoación adoptada na xeración dos enunciados, un dos aspectos da síntese vocal onde máis se está a investigar na actualidade (Fernández Rei 1999).

3.3. Procesamento documental

O ámbito do procesamento documental abrangue unha categoría moi ampla de aplicacións da lingüística computacional concibidas para a elaboración, xestión e revisión de documentos textuais. Neste apartado examinarémo-las características principais dos programas de verificación da corrección lingüística dos textos, dos programas de xeración automática de resumos, dos sistemas de extracción de información, dos sistemas de recuperación da información textual e dos programas de catalogación documental automatizada.

A revisión automática da corrección lingüística dos textos coa axuda do ordenador constitúe unha das utilidades do procesamento de textos con maior incidencia na calidade dos documentos producidos. Entre estas utilidades, as ferramentas máis utilizadas e mellor consideradas son os correctores ortográficos, mentres que os programas de verificación gramatical e estilística (moitos aínda en fase de desenvolvemento) posúen un nivel de aceptación moito menor e unha eficacia en ocasións cuestionable. A continuación, analizarémo-lo funcionamento destas utilidades de verificación lingüística automatizada, centrándonos na descrición formal dos seus obxectivos e na análise crítica das técnicas de revisión empregadas (Mitton 1996, Gómez Guinovart 1999).

Os erros de ortografía que se cometen durante a escritura dun documento co ordenador pódense orixinar por descoñecemento da norma lingüística vixente ou por distracción; os primeiros reciben a denominación técnica de *erros de competencia* e os segundos, de *erros de actuación*. Mentres que, nos erros de competencia, a causa do erro radica en que a persoa non sabe cómo se escribe a palabra, nos erros de actuación a persoa si sabe cómo se escribe a palabra pero, por algún motivo, ten un descoido ou confusión que provoca o erro.

Aínda que os erros de competencia varían moito segundo as persoas, existen determinados factores lingüísticos que favorecen a súa aparición, como a falta de correspondencia entre a ortografía e a fonética dunha palabra (pénsese, por exemplo, nas causas que contribúen o erro de **borcalladas* por *vorcalladas*), as discrepancias entre a normativa e o uso (**líbido* por *libido*) ou a interferencia con outras normativas típica das situacións de plurilingüismo (**hirmán* por *irmán*, por interferencia co castelán *hermano*). Por outra parte, os erros de actuación poden reflecti-los erros da fala (**desmolarizar* por *desmoralizar*), poden se-lo resultado dun erro mecanográfico (**escritutra* por *escritura*, debido á pulsación das dúas letras *t* e *r*, veciñas no teclado) ou poden orixinarse nunha distracción da atención (**problemente* por *probablemente*, co segmento -*lemente* colocado trala letra *b* equivocada).

Sen embargo, malia as súas diferentes causas, e con vistas ó seu tratamento informático, a maioría dos erros ortográficos poden ser descritos mediante catro mecanismos formais: inserción dunha letra (**escritutra* por *escritura*), elisión dunha letra (**hirmán* por *irmán*), substitución dunha letra por outra (**borcalladas* por *vorcalladas*) ou transposición de dúas letras adxacentes (**lingüítsica* por *lingüística*). Ademais, segundo se ten comprobado de maneira empírica, na práctica da escritura asistida por ordenador cométense moi poucos erros na primeira letra dunha palabra.

Como veremos inmediatamente, estas características formais dos erros ortográficos típicos da escritura asistida por ordenador están na base do deseño dos programas informáticos que serven para corrixilos.

Os correctores ortográficos proporcionados polos procesadores de textos tentan identifica-los erros ortográficos do documento e suxeri-la súa posible corrección. A técnica informática máis habitual para detectar estes erros consiste en compara-las palabras do documento cunha lista de palabras correctas almacenada no ordenador. Esta lista pode concibirse como un diccionario ortográfico normativo da lingua que inclúe tódalas formas flexivas das palabras, con tódalas súas formas complexas, derivadas e compostas. O corrector ortográfico indicará un erro simplemente cando unha palabra do texto non se atope nesta lista.

Con respecto á corrección, a técnica informática clásica aplicada neste caso consiste en partir da palabra que contén o erro ortográfico identificado e inverte-los devanditos catro mecanismos formais de erro. Deste xeito, cando o corrector identifica unha palabra con erro no texto, trata de buscar no diccionario as posibles formas correctas entre as palabras que comecen pola mesma primeira letra (onde xeralmente non se producen erros) e que só supoñan un tipo de erro (inserción, elisión, substitución ou transposición). Se a busca resulta infructuosa, o corrector pode amplia-lo seu ámbito de busca ás palabras que comecen por unha letra distinta (**sberta* por *aberta*) ou ás que supoñan máis dun tipo de erro (**aetra* por *aberta*, con elisión e transposición).

Estas técnicas simples de identificación das palabras ortograficamente erróneas poden fallar por diversos motivos. Ás veces, o programa corrector indica un erro onde non o hai porque a palabra buscada, a pesar de ser correcta, non está na lista de palabras utilizada polo programa. Esto acostuma suceder cos nomes propios, os neoloxismos, os tecnicismos e as palabras pouco usuais, e normalmente queda resolto mediante a ampliación individual do diccionario ortográfico normativo empregado polo procesador. Noutras ocasións, a solución non é tan sinxela, xa que o erro ortográfico cometido dá lugar a unha palabra ortograficamente correcta, diferente da pretendida, que si se atopa no diccionario do sistema (**arde* por *orde*). Se a secuencia resultante vulnera as regras sintácticas do idioma (**a arde* por *a orde*), o erro poderá ser identificado polo corrector sintáctico; en cambio, se non as vulnera, o máis fácil é que a incorrección pase desapercibida para o ordenador.

Os correctores sintácticos son os programas encargados de recoñecer e corrixi-los erros gramaticais presentes nos enunciados do documento. En comparación cos correctores ortográficos, o seu ámbito de aplicación é moito máis impreciso. Mentres que sempre se pode determinar se unha secuencia de caracteres respecta ou infrinxe as regras ortográficas instituídas pola normativa dunha lingua, non sempre é fácil para o ordenador decidir de maneira automática se unha secuencia de palabras ortograficamente correctas contén un erro sintáctico ou non, xa que as indicacións recollidas nas gramáticas normativas das linguas nunca son tan exhaustivas coma para abranguer tódolos tipos de enunciados que debe manexar un procesador de textos.

A técnica informática máis utilizada para a identificación dos erros gramaticais ten un enfoque casuístico e baséase no recoñecemento de certos patróns de erro previamente establecidos. Isto significa que o corrector sintáctico percorrerá o texto analizado tratando de detecta-las secuencias de palabras que sigan unhas determinadas pautas. Estas pautas ou patróns de erro adoitan limitarse ó nivel gráfico e poden incorporar unha suxestión de corrección. Por exemplo, un patrón simple de erro sintáctico para o galego podería ser "*che se* > *se che*". Este patrón serviríalle ó programa

corrector para identifica-la secuencia errónea **mira que che se apaga* e para suxeri-la súa substitución pola secuencia correcta *mira que se che apaga*. A técnica pode refinarse introducindo abreviaturas e símbolos nos patróns de erro. Por exemplo, o corrector pode utilizar un patrón como "DICIR *de que* > DICIR *que*" para detectar e corrixi-lo uso dequeísta do verbo *dicir* en tódalas súas formas flexivas; ou un patrón coma "*se* ... IMPSUBX ... FUTIND > *se* ... IMPSUBX ... COND" (onde IMPSUBX simboliza calquera verbo en imperfecto de subxuntivo, e os puntos suspensivos representan calquera secuencia de palabras dentro do enunciado) para identificar correlacións temporais incorrectas como **se viñeses, eu tamén irei* e para propoñe-la versión correcta *se viñeses, eu tamén iría*.

Obviamente, os resultados desta técnica dependerán da amplitude e precisión dos patróns establecidos para o programa. O corrector sintáctico só detectará un erro cando este se corresponda con algún dos patróns previstos, e non tódolos erros gramaticais son facilmente previsibles. Xa que logo, a verificación sintáctica por patróns precisa complementarse con outras técnicas de maior complexidade, como a análise sintáctica automática ou o tratamento probabilístico da coaparición léxica.

Outra ferramenta da escritura asistida por ordenador para a revisión da corrección lingüística no ámbito do procesamento de textos son os correctores estilísticos. En xeral, este tipo de correctores realiza a función de comprobar se os trazos lingüísticos do documento analizado son afíns ou non coas características atribuídas ó xénero textual ó que se adscribe o documento. Antes de que o programa corrector efectúe a revisión do documento, a persoa usuaria do sistema debe indicar a qué variedade estilística pertence o texto examinado. Deste xeito, o ordenador levará a cabo a revisión comparando as características do documento cos trazos lingüísticos establecidos como preceptivos para a categoría textual seleccionada. Normalmente, esta categoría textual pode seleccionarse a partir dun número de modelos estilísticos predefinidos polo programa. Cada un destes modelos está definido mediante un conxunto de trazos lingüísticos formais, como o número máximo de palabras por oración, a presencia ou ausencia de determinados xiros, ou o número máximo de sintagmas preposicionais consecutivos. Así mesmo, algúns correctores permiten que a persoa usuaria do sistema elabore os seus propios modelos estilísticos, asignando os valores desexados ás características lingüísticas propostas polo programa. Para que esta técnica informática de verificación estilística atinxa un grao considerable de eficiencia cómpre establecermos desde o principio cáles son as variantes estilísticas ou xéneros dunha lingua, e cáles son os trazos lingüísticos que caracterizan cada unha das variantes establecidas; dúas esixencias de difícil cumprimento ás que hai que engadi-la dificultade de que os trazos lingüísticos empregados na caracterización resulten manexables informaticamente.

Outra das técnicas máis comúns de verificación estilística na escritura asistida por ordenador consiste na avaliación do nivel de lexibilidade do texto, é dicir, do grao de dificultade de comprensión do sentido do texto determinado por certos factores lingüísticos cuantificables, como a extensión das oracións, a lonxitude das palabras ou a cantidade de preposicións dentro dunha frase. As técnicas de avaliación da lexibilidade baséanse nas regularidades estatísticas que presentan os textos neste tipo de factores, en función do seu grao de dificultade de lectura. Partindo de estudios empíricos, elabóranse mediante métodos estatísticos fórmulas ou ecuacións de lexibilidade que, combinando os valores dun conxunto de factores lingüísticos cuantificables no texto, serven para predicir parámetros estimativos do seu grao de dificultade. O procedemento clásico para

a elaboración destas fórmulas institúe que, en primeiro lugar, se estableza o conxunto de trazos lingüísticos obxectivables que hipoteticamente inciden no nivel de lexibilidade dun texto; a continuación, cómpre obter mediante enquisas os índices de dificultade dos textos dun corpus, elaborado como modelo da variedade lingüística a exame; en terceiro lugar, hase calcula-la correlación estatística existente entre os trazos estilísticos preestablecidos e os índices de dificultade obtidos empiricamente; finalmente, cos datos obtidos, hase elaborar una fórmula de lexibilidade, a modo de ecuación, cos parámetros estilísticos de maior capacidade predictiva e menor grao de correlación mutua. Por tanto, a fiabilidade que podemos outorgar ós resultados proporcionados por cada unha destas fórmulas dependerá da solidez dos seus alicerces estatísticos e da adecuación do texto analizado á variedade estilística utilizada como modelo para a elaboración da fórmula.

Deixando á marxe a revisión automática da corrección lingüística dos textos, outra das aplicacións da lingüística computacional no eido do procesamento documental é a extracción da información, que consiste en converter textos en información estructurada, por exemplo, en rexistros dunha base de datos. Así, o resultado da extracción da información dun artigo de xornal sobre unha acción terrorista podería consistir nunha ficha onde constase o tipo de incidente, a data do suceso, o lugar no que aconteceu, a identificación dos seus responsables, os obxectivos da acción, as súas consecuencias e os medios empregados, como mostra a figura 21 (adaptación ó galego simplificada dun exemplo citado en Grishman 1997).

19 de marzo.- Esta mañá un comando da guerrilla urbana salvadoreña fixo estoupar unha bomba preto dunha central eléctrica de San Salvador. O artefacto causou estragos de diversa consideración, deixando unha grande parte da poboación sen electricidade, aínda que non se rexistraron danos persoais.

TIPO DE INCIDENTE	explosión
DATA	19 de marzo
LUGAR	San Salvador
RESPONSABLES	comando da guerrilla urbana
OBXECTIVOS MATERIAIS	central eléctrica
OBXECTIVOS HUMANOS	---
DANOS MATERIAIS	estragos
DANOS PERSOAIS	non
INSTRUMENTO	bomba

Figura 21. Exemplo de extracción de información: texto e rexistro.

Xa que logo, o proceso de extracción da información supón localizar un conxunto de datos concretos no texto analizado, e construír con eles unha representación estructurada. A técnica básica utilizada para a localización dos datos consiste en identificar no texto os patróns léxico-sintácticos nos que se considera que se poden concretar lingüisticamente as informacións buscadas. Esta técnica de recoñecemento de patróns vai precedida habitualmente da anotación morfosintáctica das palabras do texto, e dunha análise sintáctica parcial centrada na identificación dos grupos nominais e verbais dos enunciados (Grishman 1997, Appelt 1999).

A diferencia da *comprensión da linguaxe natural* (Allen 1995), a extracción da información non pretende representar toda a información dun texto, senón unicamente a

información seleccionada, coa finalidade de ofrecer unha vía eficiente de consulta ós grandes volumes de datos escritos en linguaxe natural nas noticias dos xornais, nos informes médicos hospitalarios ou nas sentencias xudiciais.

Así mesmo, cómpre distinguirmos tamén a extracción da información das técnicas de *catalogación documental automatizada*, de *xeración automática de resumos* e de *recuperación da información textual*. As técnicas de catalogación documental tratan de determinar automaticamente o contido xeral dos textos analizados, para poder clasificalos dentro dunha tipoloxía semántica preestablecida. Por exemplo, un texto bancario pode ser catalogado como "débito por domiciliación" e outro como "contrato de conta". O índice bibliográfico obtido da clasificación documental constitúe unha vía de acceso simple e directa á información contida nos textos.

A xeración automática de resumos permite presenta-la información dos documentos de maneira sinóptica, o que facilita a posibilidade de realizar unha avaliación visual rápida da súa pertinencia para unha necesidade concreta de información. A técnica básica empregada para a xeración de resumos consiste na extracción das frases consideradas máis significativas do texto orixinal. As frases son seleccionadas por incluír certas palabras (por exemplo, palabras moi frecuentes no texto ou palabras que aparecen no título), ou por aparecer nun contexto determinado (por exemplo, cando a frase aparece a primeira do documento). Outras técnicas de raizame máis lingüística, aplicadas polo de agora só en dominios temáticos restrinxidos, implican a interpretación semántica do contido do documento orixinal e a posterior xeración do texto do resumo.

A recuperación da información textual é unha tecnoloxía orientada á xestión de bases de datos documentais (Strzalkowski 1999). O obxectivo é selecciona-los documentos máis relevantes en relación cuns determinados requisitos de información expresados nunha consulta. As consultas poden combina-los termos buscados mediante operadores lóxicos e condicións de proximidade. Para a selección dos documentos, utilízanse dúas listas de palabras: unha formada por tódalas palabras que aparecen nos documentos xunto coa súa localización nos textos, e outra coas palabras consideradas irrelevantes para as buscas documentais (Codina 1993). A consulta dunha base de datos textual mediante un sistema de recuperación da información ofrece como resultado a lista de documentos da base de datos que o sistema considera relevantes para satisface-la consulta. Por exemplo, o resultado de buscar na base de datos de noticias da axencia EFE as empresas que asinaron contratos relacionados coas telecomunicacións durante o pasado ano consistiría nunha lista de noticias que se debería repasar visualmente para extrae-la información desexada. En contraste, un sistema de extracción da información ofrecería directamente a lista das empresas.

3.4. Traducción automática

A traducción automática por ordenador constitúe asemade unha das aplicacións da lingüística computacional de maior complexidade intrínseca e un dos desenvolvementos de maior interese para o público non especialista. En sentido amplo, o campo da traducción automática abrangue o conxunto de ferramentas informáticas deseñadas para a súa incorporación no proceso da traducción humana. As aplicacións que forman parte deste conxunto poden agruparse segundo distintos criterios: o número de pares de linguas entre os que o sistema traduce (sistemas bilingües ou plurilingües), se traduce os pares de linguas nunha soa dirección ou tamén é quen de facer traducción

inversa (sistemas unidireccionais ou bidireccionais), se está limitado ou non no seu ámbito lingüístico a unha área temática concreta ou a un determinado tipo de lingua simplificada (traducción de linguaxe en xeral ou traducción de sublinguaxes), segundo o modelo de traducción aplicado (traducción directa, por transferencia ou mediante interlingua), ou segundo o grao de automatización da traducción (traducción automática ou traducción asistida por ordenador) (Hutchins e Somers 1992, Whitelock e Kilby 1995).

O termo *traducción automática*, en sentido estricto, refírese ós programas de traducción que non requiren intervención humana para realiza-la súa tarefa. Ata agora este tipo de aplicacións só ofrece un nivel de fiabilidade aceptable na traducción de sublinguaxes, particularmente en dominios de coñecemento moi restrinxidos (linguaxes sectoriais) ou cando o texto de partida está escrito seguindo unhas normas moi estrictas orientadas á simplificación do seu léxico e sintaxe (linguaxes controladas). Por exemplo, o programa TAUM-MÉTÉO, empregado intensivamente polo Departamento de Medio Ambiente de Canadá desde 1977, traduce os partes meteorolóxicos do inglés ó francés sen a penas necesidade de revisión humana.

Dentro da categoría da *traducción asistida por ordenador*, debe distinguirse entre a traducción semiautomática (con intervención humana) e a traducción (humana) con axuda do ordenador. Os programas informáticos de traducción semiautomática ofrecen unha traducción en borrador do texto orixinal que debe ser revisada a conciencia para acadar unha calidade de traducción similar á profesional humana. No ámbito da informática persoal, os programas deste tipo máis populares arestora son os comercializados con diversas denominacións pola empresa Globalink, mentres que para estacións de traballo o sistema máis utilizado é SYSTRAN, adoptado desde 1981 pola Comisión das Comunidades Europeas para as súas traduccións de uso interno.

Os programas de traducción con axuda do ordenador están concibidos para colaboraren como asistentes na traducción humana dun texto. Por exemplo, os contornos integrados de traballo con memoria de traducción —como TRANSLATIONMANAGER (de IBM)— integran nun único producto informático un procesador de textos especialmente deseñado para traducir, un conxunto de diccionarios bilingües, ferramentas de xestión das bases de datos léxicas e unha memoria de traducción. A memoria de traducción é unha base de datos onde se almacenan a versión orixinal e traducida de cada unha das frases que se traducen. Cando se está a traducir unha frase, o programa detecta automaticamente se esa mesma frase ou outra frase similar xa foi traducida con anterioridade, co obxecto de que se poida reutiliza-la traducción sen necesidade de reescribila completamente, facendo as modificacións que se consideren máis oportunas.

A estratexia ou modelo de traducción aplicada polos programas de traducción automática e semiautomática pode ser directa, por transferencia ou mediante interlingua. Na traducción directa, o procesamento do texto fonte produce directamente o texto traducido, sen que exista ningún tipo de análise semántica ou sintáctica intermedia. Na súa versión máis simple, pode caracterizarse como unha traducción palabra a palabra, onde a traducción se realiza substituíndo cada palabra do orixinal pola palabra correspondente no diccionario bilingüe consultado polo sistema. Como complemento adoita engadirse un certo procesamento morfolóxico das palabras do texto orixinal que permita utilizar un diccionario máis reducido de formas non flexionadas, ou algunha reorganización sinxela das categorías léxicas do texto da traducción (por

exemplo, na traducción entre galego e inglés, inverte-la orde das palabras cando se identifique unha secuencia formada por un nome seguido dun adxectivo) (Figura 22).

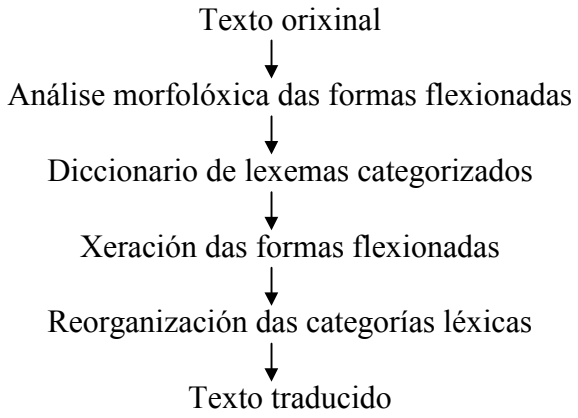


Figura 22. Estratexia de traducción directa ampliada.

Na traducción mediante interlingua, analízase o texto orixinal para construír unha representación intermedia a partir da que se xera directamente o texto da traducción. A representación intermedia é unha formalización do significado conceptual do texto, polo que constitúe unha representación abstracta tanto do texto orixinal coma do texto traducido. O termo *interlingua* alude a que esta representación semántica pretende ser neutral respecto ás linguas traducidas e, nalgunhas das súas formulacións, mesmo universal (é dicir, neutral respecto a tódalas linguas naturais). A estratexia é moi axeitada para os sistemas plurilingües xa que, unha vez definida a interlingua, só esixe dous módulos de programación para cada lingua L incorporada bidireccionalmente ó sistema: un módulo de análise (ou traducción de L á interlingua) e un de xeración (ou traducción de interlingua á L) (Figura 23).

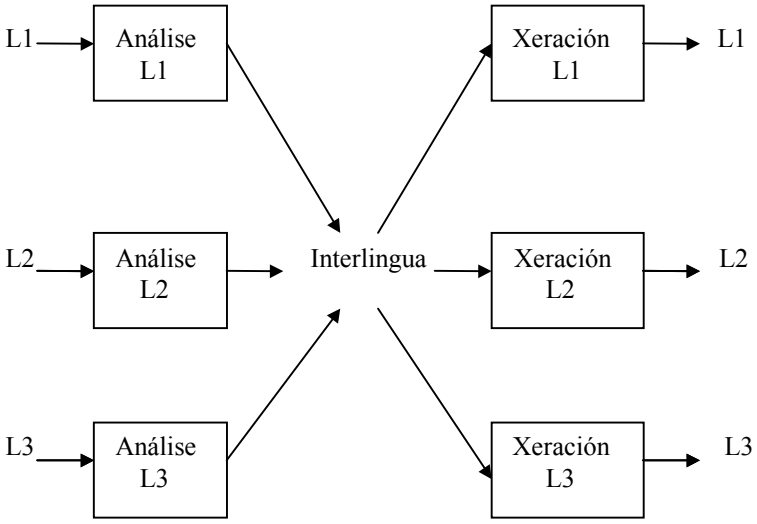


Figura 23. Estratexia de traducción mediante interlingua.

No método de traducción por transferencia, a traducción realízase en tres fases: a fase de análise do texto orixinal, que produce como resultado unha representación sintáctico-semántica dependente da lingua analizada; a fase de transferencia, na que se substitúen as palabras da lingua do orixinal polas palabras da lingua da traducción e se converte a estructura sintáctico-semántica propia da lingua orixinal nunha estructura equivalente na lingua da traducción; e a fase de xeración, na que se transforma a representación sintáctico-semántica froito da transferencia nun texto na lingua da traducción. Con respecto ó modelo da interlingua, esta estratexia ten a vantaxe de que os seus módulos de análise e xeración son menos complexos, xa que traballan con representacións moi próximas ás linguas representadas. Por contra, o seu principal defecto é o número de módulos necesarios para construír un sistema plurilingüe: para traducir bidireccionalmente entre n linguas, cómpre elaborar un total de $n^2 + n$ módulos, fronte ós $2 \times n$ módulos necesarios na aproximación interlingüística. Por exemplo, nun sistema trilingüe por transferencia os módulos ascenden a 12 (6 de transferencia, 3 de análise e 3 de xeración), fronte ós 6 esixidos nun sistema de interlingua (3 de análise e 3 de xeración) (Figura 24).

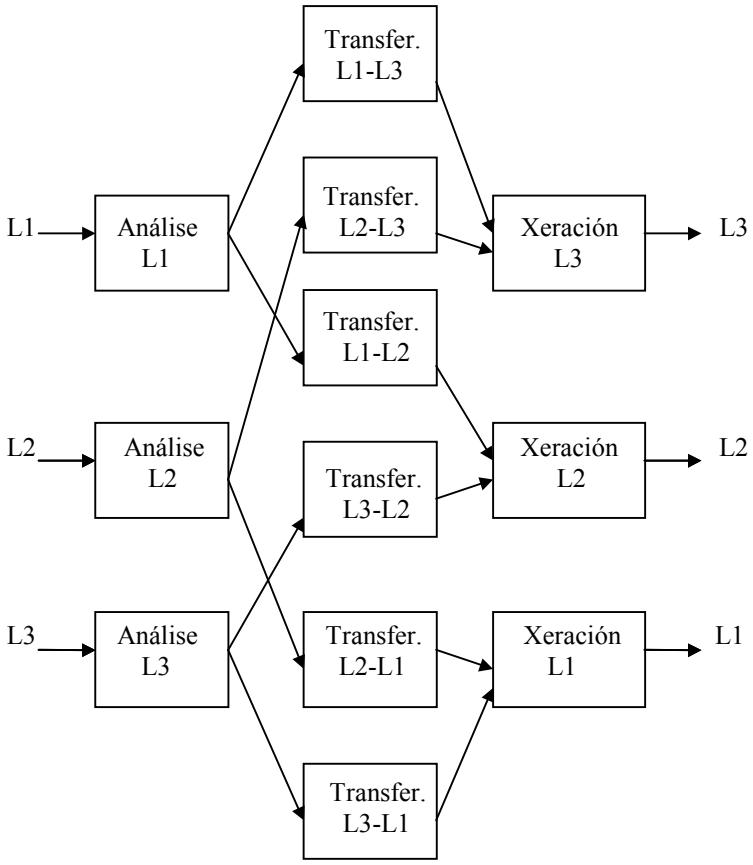


Figura 24. Estratexia de traducción por transferencia.

4. INFORMÁTICA APLICADA Á LINGÜÍSTICA

A denominación de *informática aplicada á lingüística* (ou simplemente *lingüística informática*) abarca, na súa acepción máis xeral, toda a variedade de estudios lingüísticos que utilizan ferramentas informáticas para a súa investigación e, de maneira máis específica, aqueles onde a aplicación da informática posúe unha maior influencia na obtención dos seus resultados. Nesta alínea examinaremos de forma resumida dous exemplos ilustrativos da investigación neste campo, centrando a nosa atención na lingüística comparada diacrónica e na lingüística de corpus. Outras liñas de investigación salientables da lingüística informática son a informática aplicada á lingüística antropolóxica (Antworth e Valentine 1998), a informática aplicada á lexicografía (Boguraev e Briscoe 1989, Ooi 1998, Martí, Castellón e Fernández 1998, Álvarez Lugrís 1999, Pérez, Moreno e Faber, 1999) e a informática aplicada á sociolingüística (Moreno Fernández 1994, Lorenzo 1999, Ramallo 1999).

4.1. Lingüística de corpus

A lingüística de corpus é unha disciplina dedicada ó estudio empírico da linguaxe a partir dos datos que proporcionan os corpus lingüísticos (Badia 1996, Llisterri 1996, McEnery e Wilson 1996, Pérez Guerra 1998). Un corpus lingüístico consiste nunha colección de textos reais escritos ou falados, almacenados en soporte informático, compilados para os efectos dalgunha investigación ou aplicación lingüística, e representativos dunha variedade lingüística determinada (por exemplo, do inglés escrito do século XV ou dos noticiarios radiofónicos emitidos en galego no período comprendido entre 1975 e 1995).

A selección dos textos que han formar parte dun corpus vai depender dos seus obxectivos. Os corpus xerais pretenden abranguer tódalas variedades e rexistros dunha lingua, co fin de acadaren a máxima representatividade lingüística, mentres que os corpus especializados (monolingües ou plurilingües) prefiren cinguirse a unha variedade lingüística concreta (por exemplo, o *Corpus Legebiduna* inclúe textos administrativos e xurídicos bilingües publicados en lingua vasca e castelá polas institucións do País Vasco) (Abaitua, Casillas e Martínez 1997). Cando os textos plurilingües recompilados son as versións traducidas dos mesmos documentos fálase de corpus de *textos paralelos*; nos corpus de *textos aliñados*, ademais, están identificadas as equivalencias de traducción entre os segmentos (palabras, frases ou unidades de traducción) de cada unha das versións traducidas (Hallebeek 1999).

Os corpus *crus*, é dicir, os corpus que conteñen exclusivamente as palabras e signos de puntuación dos textos orixinais, teñen unha utilidade bastante limitada para a investigación lingüística de corpus. Así, nun corpus cru sería complicado localizar automaticamente (e sen ningunha análise previa) os sintagmas nominais, os tonemas descendentes ou as oracións nas que o suxeito non está en posición inicial. Para facilitaren estas buscas e outras semellantes, os corpus *anotados* acompañan o texto dos documentos orixinais de diversos tipos de información lingüística elaborada de maneira manual, semiautomática ou completamente automática.

Na actualidade, a anotación morfosintáctica —é dicir, o proceso de asignar unha indicación da súa categoría gramatical a cada palabra dun texto— pode efectuarse cun alto grao de automatización. Os programas informáticos de marcaxe gramatical automática analizan as secuencias de palabras para produciren as etiquetas

morfosintácticas correspondentes. A lista de etiquetas dependerá das características lingüísticas do corpus, dos obxectivos da súa anotación, dos límites do programa e dos presupostos teóricos aplicados (Garside, Leech e McEnery 1997). A figura 25 mostra un exemplo de anotación gramatical real (lixeiramente modificada) do *Corpus LOB*, co texto cru orixinal e o texto producto da marcaxe (*apud* McEnery e Wilson 1996, 37).

```
Joanna stubbed out her cigarette with unnecessary fierceness.
Her lovely eyes were defiant above cheeks whose colour had
deepened at Noreen's remark.

Joanna_NP stubbed_VBD out_RP her_PP$ cigarette_NN with_IN
unnecessary_JJ fierceness_NN ._. Her_PP$ lovely_JJ eyes_NNS
were_BED defiant_JJ above_IN cheeks_NNS whose_WP$ colour_NN
had_HVD deepened_VBN at_IN Noreen's_NP$ remark_NN ._.
```

Figura 25. Exemplo de marcaxe morfosintáctica: texto cru e anotado.

Os etiquetadores morfosintácticos automáticos actuais poden utilizar regras lingüísticas ou modelos probabilísticos como mecanismo de asignación de etiquetas. Tamén hai sistemas etiquetadores híbridos que combinan estas dúas estratexias en función das súas necesidades. A maioría dos etiquetadores probabilísticos —os máis estudiados e difundidos— asignan as etiquetas utilizando a información morfolóxica e contextual de tipo estatístico derivada automaticamente dun corpus inicial amplo, etiquetado previamente, que constitúe o modelo de actuación lingüística no que se basearán as análises morfosintácticas efectuadas polo programa. Así, o etiquetador fundamentará a súa actuación en dúas estimacións estatísticas: a *probabilidade léxica*, baseada na frecuencia relativa coa que unha palabra recibiu unha etiqueta determinada no corpus inicial; e a *probabilidade contextual*, baseada na frecuencia relativa coa que unha determinada etiqueta aparece antes das *n* seguintes etiquetas e/ou despois das *n* anteriores etiquetas no corpus inicial. O valor de *n* ha depender do deseño concreto do etiquetador, aínda que parece haber consenso en que a súa maior eficiencia se obtén asignándolle á variable *n* o valor de 1 ou de 2.

A marcaxe das palabras descoñecidas (é dicir, as que non constan no corpus inicial) é un dos principais problemas para os etiquetadores probabilísticos. A solución máis habitual consiste en etiqueta-la palabra descoñecida en función do seu contexto (de acordo co modelo probabilístico derivado do corpus inicial empregado polo etiquetador), ponderando este factor contextual con outras claves deducibles da grafía da palabra. Por exemplo, a presencia dunha maiúscula inicial aumentaría a probabilidade de asignación da etiqueta correspondente ós nomes propios, e o seu final en *-ción* incrementaría as súas posibilidades de recibi-la etiqueta destinada ós substantivos deverbais.

Os corpus anotados morfosintacticamente son moi útiles para a investigación lingüística e para o desenvolvemento de aplicacións de procesamento da linguaxe natural, xa que proporcionan textos con palabras non ambiguas respecto á súa categoría gramatical. Así, nun corpus de galego con anotación morfosintáctica, sería doado localiza-las perífrases de pasiva, os pronomes preverbais ou os casos de artigo seguido de adxectivo, por mencionarmos tres exemplos ilustrativos do seu uso nos estudios lingüísticos.

Certamente, un corpus anotado con información sintáctica —é dicir, un corpus analizado sintacticamente e etiquetado con esta información— resulta aínda máis valioso. Nembargantes, os corpus amplos anotados sintacticamente son escasos, xa que deben ser elaborados de maneira manual ou semiautomática por mor das dificultades da automatización completa do proceso da análise sintáctica. Véxase na figura 26 un exemplo de marcaxe sintáctica do *Lancaster Parsed Corpus*, co texto orixinal e a continuación o texto etiquetado (*apud* Pérez Guerra 1998, 37).

```
I can n't make a club pay a player so much a week.

[S[Na I_PP1A Na] [V can_MD n't_XNOT make_VB V][N a_AT club_MM
N][Tb[V pay_VB V] [N a_AT player_NN N][N[D so_QL much_AP D][N
a_AT week_NN N]N]Tb]._. S]
```

Figura 26. Exemplo de marcaxe sintáctica.

Outros tipos de anotación lingüística que se poden atopar ocasionalmente nos corpus dispoñibles na actualidade son a anotación prosódica (indicacións sobre pausas na pronuncia, curvas de entoación, grupos fónicos, indicacións de énfase, etcétera), a anotación semántica (indicacións sobre características semánticas dos elementos léxicos ou sobre a súa pertenza a un campo semántico), a anotación discursiva (como indicacións sobre a referencia dos pronomes) e a anotación dos fenómenos propios da conversación (por exemplo, indicacións sobre as quendas de fala o sobre os elementos fáticos).

En canto ó formato de marcaxe, existen moitas maneiras de representa-la información lingüística nos textos. A única norma de representación xeralmente respectada é que tódalas etiquetas empregadas se poidan suprimir con facilidade. As anotacións da figura 1, por exemplo, poden ser eliminadas con pouco esforzo borrando as secuencias de caracteres situados entre un guión e un espacio.

Outro formato de anotación cada vez máis habitual na marcaxe de corpus é o formato SGML (sigla de *Standard Generalised Markup Language*) e, en particular, un subconxunto de SGML coñecido como formato TEI (*Text Encoding Initiative*) (Sperberg-McQueen e Burnard 1994). Sen entrarmos en moitos detalles, xa que unha presentación exhaustiva das normas TEI excedería os límites deste capítulo, cómpre saber que estas directrices proporcionan un conxunto normalizado de etiquetas e abreviaturas para a representación descritiva e estructurada da información textual contida nos corpus. As etiquetas TEI deben escribirse entre parénteses triangulares (como en <*etiqueta*>) e poden ser dobres para delimitaren unha porción de texto entre <*etiqueta*> e <*/etiqueta*>, mentres que as abreviaturas (codificadas como *&abreviatura;*) permiten condensa-la información incorporada no texto e representa-los caracteres e signos gráficos con dificultades de traducción entre os distintos sistemas informáticos. Os documentos en formato TEI divídense en dúas partes: a súa cabeceira e o propio texto anotado. A cabeceira contén as informacións bibliográficas e de codificación relativas ó documento electrónico: autoría, data de publicación, título, edición, procedencia da versión electrónica, observacións sobre o formato de anotación utilizado, etcétera. Por outra banda, o texto estructúrase en tres seccións: os preliminares (datos da primeira páxina, prefacio, limiar, dedicatoria, índice, etc.), o corpo (subdividido xerarquicamente en diversos elementos) e a parte final (apéndices,

notas, bibliografía, glosario, etc.). A modo de ilustración, a figura 27 contén un exemplo de anotación textual con TEI dun fragmento do conto "Tinta chinesa" de Gonzalo Navaza.

```
<tei.2>
<teiheader> <filedesc> <titlestmt> <title> Fragmentos do
conto "Tinta chinesa", de Gonzalo Navaza: un exemplo de
codificaci&oacute;n TEI </title> <respstmt> <resp> preparado
por </resp> <name>Xavier G&oacute;mez Guinovart </name>
</respstmt> </titlestmt> <publicationstmt> <publisher>
Edici&oacute;ns Xerais de Galicia </publisher>
</publicationstmt> <sourcedesc> <bibl> <author> Gonzalo
Navaza </author> <title type="item"> Tinta chinesa </title>
<title> Erros e T&aacute;natos </title> <imprint> <publisher>
Edici&oacute;ns Xerais de Galicia </publisher> <pubplace>
Vigo </pubplace> <date> 1996 </date> </imprint> <extent> pp.
43-56 </extent> </bibl> </sourcedesc> </filedesc>
</teiheader>

<text> <front> <titlepage> <doctitle> <titlepart> Tinta
chinesa </titlepart> </doctitle> </titlepage> <div1
type="dedication"> <pb>  <p> A Arthur M. Morgan, <foreign> in
memoriam </foreign> . </p> </div1> </front>

<body> <p> En <date value="09-1993"> setembro de 1993
</date>, durante os preparativos da celebraci&oacute;n do
congreso do <foreign>PEN Club</foreign> Internacional en
Santiago de Compostela, os coitados mozos e menos mozos
aspirantes a ingresarmos en tan ilustre asociaci&oacute;n de
escritores recibimos da organizaci&oacute;n a encomenda de
recoller no aeroporto alg&uacute;ns dos asistentes e
conducilos ata o hotel. Trat&aacute;base en xeral de
escritores de segunda fila, pois os pesos pesados &iacute;an
ser atendidos directamente polos organizadores. O encargo
inclu&iacute;a a posibilidade de facer de cicerone polas
r&uacute;as e monumentos da cidade e obrigaba a estar sempre
disposto para resolver calquera problema que puidese
present&aacute;rselles &oacute;s convidados. </p> <note
resp="XGG"> texto elidido (dous par&aacute;grafos) </note>
<pb n="46"> <p> Daquela despedida quero evocar tam&eacute;n,
se se me permite, outro detalle. Cando xa se retiraran os
fot&oacute;grafos e o resto dos acompa&ntilde;antes,
m&iacute;ster Morgan coleunos &aacute; parte a Marcelo
Cardalda e mais a min e, obrig&aacute;ndonos a coloca-las
palmas das mans coma no xuramento dos tres mosqueteiros,
berrou en franc&eacute;s <q rend="<< >>"> <foreign> Un pour
tous, tous pour un! </foreign> </q> , coa s&uacute;a voz de
b&uacute;falo e o seu curioso acento. Logo deunos un abrazo
efusivo e antes de perderse polo corredor de embarque
entregounos un pequeno obsequio: a Carralda <title
rend="italic"> <foreign> A Touch of Class </foreign> </title>
, nunha edici&oacute;n ilustrada por Bacon, e a min os <title
rend="italic"> <foreign> Forgettable Tales </foreign>
</title> , na edici&oacute;n de Planet, encadernada en pel.
</p> <note resp="XGG"> resto de texto elidido </note> </body>
</text> </tei.2>
```

Figura 27. Exemplo de marcaxe TEI.

4.2. Lingüística histórica computacional

A lingüística histórica estudia os cambios que se producen nas linguas nunha dimensión temporal ou diacrónica. Un dos seus obxectivos consiste en reconstruír deductivamente os estados hipotéticos anteriores dunha lingua, a partir das formas lingüísticas posteriores testemuñadas. Cando estas formas pertencen a diferentes variedades lingüísticas, presumiblemente relacionadas entre si, a reconstrucción da protolingua —é dicir, da lingua común hipotética da que se orixinaron— procede segundo o denominado *método comparativo*. O proceso de reconstrucción lingüística comeza coa identificación dos conxuntos de palabras derivadas dun mesmo étimo nas diversas linguas (como o inglés *father*, o grego *pater* e o sánscrito *piter*) e, a partir destes grupos de *cognados*, reconstrúense os étimos (por exemplo, **pAter*) e formúlanse as regras diacrónicas que describen os cambios fonéticos observados (en inglés, ***p* > *f*).

O programa PHONO[4] permite comproba-los efectos das teorías postuladas, sempre que se lle forneza dun conxunto ordenado de regras fonolóxicas e dunha descrición dos trazos distintivos das vocais e consoantes que compoñen o alfabeto. A partir dun étimo da protolingua, PHONO xerará automaticamente a cadea derivativa prevista na teoría, manipulando as matrices de trazos de acordo coas regras establecidas (Figura 28).

```
ETYMON -->                      mensa
  HOMORGANIC:              =>   mén Sa   -ante-dist
  PRENASAL_LONG:           =>   mén Sa   +long/-ante-dist
  NS_S:                    =>   mé Sa    +long
  UNLONG:                  =>   méSa
  VOICING:                 =>   méZa
  UNVOICE:                 =>   méSa
```

Figura 28. Derivación de *mensa* do latín ó castelán con PHONO.

Outros programas informáticos, como COGNATE[5] (Guy 1994) e WORDSURV[6], ofrecen a posibilidade de calcula-lo grao de parentesco entre palabras de diferentes linguas, baseándose exclusivamente na probabilidade estatística das súas correspondencias fonéticas (Figura 29).

```
    a   p   p   l   e
a  93  32  32  78  68
p  21  83  83  66  14
f  34  86  86  52  70
e  49  63  63  30  91
l  44  54  54  97  41

apfel/apple  (word pair #3, pass #1)
I am 39% sure that they ARE related.
I allowed only for matches >= 50.
The best I found were: apfel
                       app le
```

Figura 29. Probabilidades de cognación entre *apple* e *apfle* con COGNATE.

[4] http://www.siu.edu/~nmc/phono33.zip.
[5] ftp://garbo.uwasa.fi/pc/linguistics/cognate.zip.
[6] ftp://ftp.sil.org/software/dos/wrdsrv25.zip.

Finalmente, existen utilidades como RECONSTRUCTION ENGINE (Lowe e Mazaudon 1994), para a reconstrucción das formas dunha protolingua a partir dos datos lingüísticos de calquera grupo de linguas do mundo, e como GLOTTO[7], capaz de elaborar automaticamente árbores xenealóxicas de familias lingüísticas a partir de listas de palabras e da súa análise léxico-estatística. Na figura 30 ofrécese un exemplo dos resultados de GLOTTO, a partir de datos de oito linguas austronésicas de Vanuatu, onde as cantidades expresan a proporción de vocabulario (por cada mil palabras) que se mantén en cada nova división da familia (por exemplo, o sakao e o fortsenal manterían respectivamente o 56,7% e o 75,9% do léxico do seu antepasado común que, pola súa banda, conservaría un 88,3% da súa lingua antecesora).

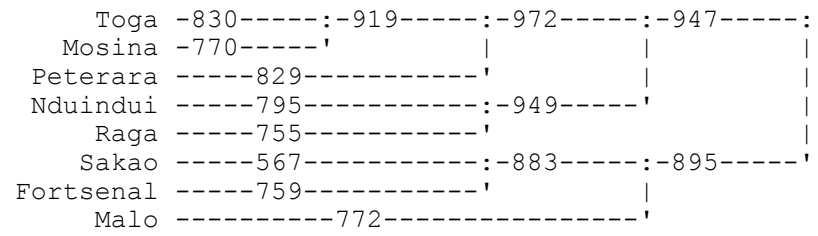


Figura 30. Árbore xenealóxica léxico-estatística elaborada por GLOTTO.

5. CONCLUSIÓNS

Neste capítulo procuramos presentar, desde unha perspectiva didáctica e global, algunhas das liñas de traballo máis salientables da lingüística computacional. Con este obxectivo, na primeira parte do traballo, tratamos de delimita-lo campo de estudio desta disciplina e de establece-las súas relacións coas outras áreas da lingüística e coa sociedade. A continuación, no segundo apartado, considerámo-lo campo de traballo da lingüística computacional teórica, atendendo particularmente ós modelos lingüísticos simbólicos e ós formalismos lingüísticos de unificación. Na terceira alínea do capítulo, presentámo-lo estado da cuestión das tecnoloxías da lingua nos eidos da comprensión e xeración da linguaxe, do recoñecemento e a síntese da fala, da extracción da información textual e da traducción automática. Por último, na sección final, examinamos algúns empregos da informática na investigación lingüística, centrándonos na súa aplicación á lingüística histórica e, de maneira especial, á análise textual de corpus lingüísticos. Mediante a lectura atenta deste capítulo e máis das referencias bibliográficas propostas, as persoas interesadas poderán obter unha visión panorámica xeral dun dos campos da lingüística con máis posibilidades de investigación e desenvolvemento e con maior incidencia social.

6. EXERCICIOS

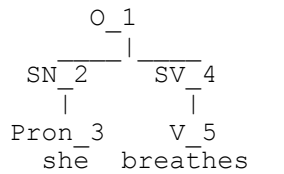
1. [1.1] Sinalar algunhas aplicacións da lingüística computacional que poidan servir para mellora-la calidade de vida das persoas con discapacidades motoras ou sensoriais.

[7] ftp://garbo.uwasa.fi/pc/linguistics/glotto02.zip.

2. [1.2] Describi-las relacións existentes entre a intelixencia artificial, a lingüística computacional, a lingüística cognitiva e a psicolingüística.

3. [2.1] Enumerar algunhas das vantaxes teóricas e descritivas que supón para unha gramática incorpora-la distinción postulada pola GPSG e a HPSG entre regras de dominio inmediato e regras de precedencia lineal. Suxestión: considera-lo fenómeno da orde libre de constituíntes na lingua vasca. Claves bibliográficas: Borsley 1996, 44-64; Ruiz 1996, 45-50.

4. [2.2] Escribir no formalismo gramatical empregado por PC-PATR as regras sintácticas e o léxico necesarios para representa-los enunciados (a, b) e bloquea-los enunciados agramaticais de (c, d): (a) *she breathes*; (b) *they breathe*; (c) **she breathe*; (d) **they breathes*. Clave orientativa: a figura 31 amosa un resultado posible de analiza-lo enunciado (a) en PC-PATR cun léxico e un conxunto de regras sintácticas axeitadas como resposta.




```
O_1:
[ cat:   O ]
SN_2:
[ cat:   SN
  num:   sg ]
Pron_3:
[ cat:   Pron
  lex:   she
  num:   sg ]
SV_4:
[ cat:   SV
  num:   sg ]
V_5:
[ cat:   V
  lex:   breathes
  num:   sg ]
```

Figura 31. Análise sintáctica de *she breathes* en PC-PATR.

5. [2.2] Representar no formalismo DATR a morfoloxía dun paradigma flexivo simple, por exemplo, a flexión regular de xénero e número dos adxectivos en galego. Clave bibliográfica: Evans e Gazdar 1996.

6. [3.3] A fórmula de lexibilidade máis utilizada para o inglés é a proposta por Flesh (1948), que determina o índice de lexibilidade (IL) dun texto a partir da media de sílabas por cada 100 palabras (SP) e a media de palabras por oración (PO), segundo o seguinte cálculo: *IL = 206,835 – (0,846 x SP) – (1,015 x PO)*. Os resultados deben interpretarse mediante a seguinte táboa de equivalencias (Figura 32), que se acompaña

dos patróns estilísticos regulares respecto ós valores de SP e PO en cada grao de dificultade. Considerar cál sería o resultado de aplica-la fórmula de Flesh a textos escritos en lingua galega e explica-las razóns das posibles diferencias. Clave bibliográfica: Gómez Guinovart 1999.

IL	*Grao de dificultade*	*SP*	*PO*
0-30	Moi difícil	192 ou máis	29 ou máis
30-50	Difícil	167	25
50-60	Máis ben difícil	155	21
60-70	Estándar	147	17
70-80	Máis ben fácil	139	14
80-90	Fácil	131	11
90-100	Moi fácil	123 ou menos	8 ou menos

Figura 32. Índice de lexibilidade de Flesh para o inglés.

7. [4.1] Enumerar algunhas aplicacións concretas da análise de corpus ó estudio da linguaxe, indicando o tipo de anotación necesaria en cada caso. Exemplo: para determina-los diferentes contextos sintácticos nos que pode aparecer unha determinada palabra nos textos, cómpre dispoñer dun corpus anotado gramaticalmente. Clave bibliográfica: McEnery e Wilson 1996, 87-116.

8. [4.2] Utilizando o programa COGNATE, comproba-lo grao hipotético de parentesco léxico entre os datos lingüísticos da figura 33, calculado a partir da probabilidade estatística das correspondencias fonéticas entre as palabras de cada lingua. Clave bibliográfica: Guy 1994.

	Danés	*Vasco*	*Finés*	*Sánscrito*	*Bretón*	*Romanés*
1	en	bat	yksi	eka	unan	unu
2	to	bi	kaksi	dva	daou	doi
3	tre	hiru	kolme	tri	tri	trei
4	fire	lau	neljä	catur	pevar	patru
5	fem	bost	viisi	pañca	pemp	cinci
6	seks	sei	kuusi	ṣaṣ	c'hwec'h	ṣase
7	syu	zazpi	seitsemän	sapta	seizh	ṣapte
8	otte	zortzi	kahdeksan	aṣṭa	eizh	opt
9	ni	bederatzi	yhdeksan	nava	nav	nouă
10	ti	hamar	kymmenen	daśa	dek	zece

Figura 33. Números cardinais (1-10) en diferentes linguas.

7. BIBLIOGRAFÍA COMENTADA

Allen, James (1995), *Natural Language Understanding*, 2ª ed. Redwood: Benjamin/Cummings.

Monografía académica dedicada ó estudio dos sistemas de comprensión da linguaxe. Analiza con profundidade, desde unha perspectiva computacional, os problemas sintácticos, semánticos e pragmáticos implicados na conversión dun texto nunha representación do seu significado.

Borsley, Robert (1996), *Modern Phrase Structure Grammar*. Cambridge: Blackwell.

Introducción á gramática sintagmática xeneralizada (GPSG) e á gramática sintagmática dirixida polo núcleo (HPSG), dúas das teorías lingüísticas contemporáneas de maior implantación no ámbito da lingüística computacional. Contén exercicios prácticos de afianzamento dos conceptos desenvolvidos en cada capítulo.

Dutoit, Thierry (1997), *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers.

Manual sobre os sistemas de conversión de texto a voz, dirixido a estudiantes desta materia en lingüística computacional e enxeñería de telecomunicacións. Presenta as técnicas de procesamento da linguaxe natural e de procesamento do sinal sonoro empregadas actualmente na síntese da fala.

Gazdar, Gerald e Mellish, Chris (1989), *Natural Language Processing in Prolog: An Introduction to Computational Linguistics*. Wokingham: Addison-Wesley.

Manual de lingüística computacional concibido como libro de texto de orientación práctica. Contén unha extensa selección de exemplos e exercicios de programación en Prolog que ilustran os conceptos e técnicas de procesamento sintáctico, semántico e pragmático expostos no texto.

Grishman, Ralph (1986), *Computational Linguistics: An Introduction*. Cambridge: Cambridge University Press. (Trad. cast.: *Introducción a la lingüística computacional*, Visor, Madrid, 1991.)

Obra clásica introductoria ó eido da lingüística computacional. A súa lectura é recomendable para un primeiro achegamento a esta disciplina, malia as limitacións derivadas da falta de actualización dos seus contidos.

Hutchins, John e Somers, Harold (1992), *An Introduction to Machine Translation*. Londres: Academic Press. (Trad. cast.: *Introducción a la traducción automática*, Visor, Madrid, 1995.)

Completo manual de traducción automática. Inclúe unha ampla introducción ós fundamentos lingüísticos e computacionais da traducción automática, e unha descrición minuciosa das características de deseño de diferentes sistemas.

McEnery, Tony e Wilson, Andrew (1999), *Corpus Linguistics*, 2ª ed. Edimburgo: Edinburgh University Press.

Manual sobre as técnicas e fundamentos básicos da investigación lingüística de corpus textuais asistida por ordenador. Está concibido como un libro de texto con exemplos, exercicios e lecturas.

Shieber, Stuart (1986), *An Introduction to Unification-Based Approaches to Grammar*. Stanford: CSLI. (Trad. cast.: *Introducción a los formalismos gramaticales de unificación*, Teide, Barcelona, 1989.)

Introducción concisa a algúns dos modelos lingüísticos (LFG, GPSG, HPSG) e formalismos (PATR, DCG, FUG) máis espallados que traballan con estructuras de trazos e unificación. O libro presta particular atención á explicación dos conceptos lingüísticos e computacionais básicos, e á súa formalización en PATR.

Varile, Giovanni e Zampolli, Antonio (eds.) (1997), *Survey of the State of the Art in Human Language Technology*. Cambridge: Cambridge University Press.

Presentación polifónica exhaustiva dos métodos e aplicacións da lingüística computacional contemporánea, con especial atención ás vertentes máis relacionadas coas industrias da lingua.

Whitelock, Peter e Kilby, Kieran (1995), *Linguistic and Computational Techniques in Machine Translation System Design*. Londres: UCL Press.

Análise, comparación e avaliación crítica de seis destacados sistemas de traducción automática. Na introducción, os autores do estudio discuten diversos aspectos relacionados cos problemas lingüísticos da traducción automática, e presentan as técnicas básicas para o seu tratamento informático.

8. BIBLIOGRAFÍA COMPLEMENTARIA

Abaitua, Joseba; Casillas, Arantza e Martínez, Raquel (1997), "Segmentación de corpus paralelos para memorias de traducción". *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 21, 17-30.

Alonge, Antonietta; Calzolari, Nicoletta; Vossen, Piek; Bloksma, Laura; Castellón, Irene; Martí, M. Antonia e Peters, Wim (1998), "The Linguistic Design of the EuroWordNet Database". *Computers and the Humanities*, 32, 91-115.

Álvarez Lugrís, Alberto (1999), "Técnicas de representación en la lexicografía plurilingüe". En: Gómez Guinovart, Javier et al. (eds.), 215-245.

Antworth, Evan e Valentine, Randolph (1998), "Software for Doing Field Linguistics". En: Lawler, John e Dry, Helen Aristar (eds.) (1998), *Using Computers in Linguistics: A Practical Guide*. Londres: Routledge, 170-196.

Appelt, Douglas (1999), "Introduction to Information Extraction". *AI Communications*, 12, 161-172.

Badia, Toni (1996), "El processament computacional de corpus: tècniques automàtiques d'anàlisi morfològica i sintàctica". En: Payrató, Lluís et al., 217-254.

Balari Ravera, Sergi (1999), "Formalismos gramaticales de unificación y procesamiento basado en restricciones". En: Gómez Guinovart et al. (eds.), 117-151.

Bennett, Paul (1995), *A Course in Generalized Phrase Structure Grammar*. Londres: UCL Press.

Bird, Steven (1995), *Computational Phonology: A Constraint-Based Approach*. Cambridge: Cambridge University Press.

Boguraev, Bran e Briscoe, Ted (1989), *Computational Lexicography for Natural Language Processing*. Londres: Longman.

Bresnan, Joan (1999), *Lexical-Functional Syntax*. Oxford: Blackwell.
Carpenter, Bob e Penn, Gerald (1997), *ALE: The Attribute Logic Engine*. Pittsburgh: Universidade Carnegie Mellon.
Charniak, Eugene (1993), *Statistical Language Learning*. Cambridge: The MIT Press.
Codina Bonilla, Lluís (1993), *Sistemes d'informació documental*. Barcelona: Pòrtic.
Comisión Europea (1998), *The Euromap Report: Challenge and Opportunity for Europe's Information Society*. Luxemburgo: Linglink.
Evans, Roger e Gazdar, Gerald (1996), "DATR: A Language for Lexical Knowledge Representation". *Computational Linguistics*, 22, 167-216.
Fernández Rei, Elisa (1999), "Tecnologías del habla y síntesis de voz en gallego". En: Gómez Guinovart, Javier et al. (eds.), 103-116.
Garside, Roger; Leech, Geoffrey e McEnery, Tony (eds.) (1997), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Londres: Longman.
Gazdar, Gerald; Klein, Ewan; Pullum, Geoffrey e Sag, Ivan (1985), *Generalized Phrase Structure Grammar*. Oxford: Blackwell.
Gómez Guinovart, Javier (1999), *La escritura asistida por ordenador: problemas de sintaxis y de estilo*. Vigo: Universidade de Vigo (Servicio de Publicacións).
Gómez Guinovart, Javier; Lorenzo Suárez, Anxo M.; Pérez Guerra, Javier e Álvarez Lugrís, Alberto (eds.) (1999), *Panorama de la investigación en lingüística informática*. Monografía de *Revista Española de Lingüística Aplicada*.
Grishman, Ralph (1997), "Information Extraction: Techniques and Challenges". En: Pazienza, Teresa (ed.), *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. Berlín: Springer-Verlag, 10-27.
Guy, Jacques (1994), "An Algorithm for Identifying Cognates in Bilingual Word-lists and its Applicability to Machine Translation". *Journal of Quantitative Linguistics*, 1, 35-42.
Hallebeek, Jos (1999), "El corpus paralelo". *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 24, 49-55.
Kay, Martin (1982), "Parsing in Functional Unification Grammar". En: Dowty, David; Kartunnen, Lauri e Zwicky, Arnold (eds.), *Natural Language Parsing*. Cambridge: Cambridge University Press, 251-278.
Klavans, Judith (1997), "Computational Linguistics". En: O'Grady, William, Dobrovolsky, Michael e Katamba, Francis (eds.), *Contemporary Linguistics: an Introduction*. Londres: Longman, 664-702.
Koskenniemi, Kimmo (1983), *Two-level morphology: A general computational model for word-form recognition and production*. Helsinki: Universidade de Helsinki.
Llamas, César e Cardeñoso, Valentín (1997), *Reconocimiento automático del habla: técnicas y aplicación*. Valladolid: Universidade de Valladolid.
Llisterri, Joaquim (1996), "Els corpus lingüístics orals". En: Payrató, Lluís et al., 27-70.
Lorenzo Suárez, Anxo M. (1999), "Sociolingüística cualitativa y lingüística informática". En: Gómez Guinovart et al. (eds.), 247-261.
Lowe, John e Mazaudon, Martine (1994), "The Reconstruction Engine: A Computer Implementation of the Comparative Method". *Computational Linguistics*, 20, 381-417.
Martí, M. Antonia; Castellón, Irene e Fernández, Ana (1998), "Extracción de información de corpus diccionariales". En: Gómez Guinovart, Javier e Palomar, Manuel (eds.), *Lengua y tecnologías de la información*. Monografía de *Novática: Revista de la Asociación de Técnicos de Informática*, 133, 4-10.

Mitton, Roger (1996), *English Spelling and the Computer*. Londres: Longman.
Moreno, Lidia; Palomar, Manuel; Molina, Antonio e Ferrández, Antonio (1999), *Introducción al procesamiento del lenguaje natural*. Alicante: Universidade de Alicante.
Moreno Fernández, Francisco (1994), "Status quaestionis: sociolingüística, estadística e informática". *Lingüística*, 6, 95-154.
Moreno Sandoval, Antonio (1998), *Lingüística computacional*. Madrid: Síntesis.
Moure, Teresa e Llisterri, Joaquim (1996), "Lenguaje y nuevas tecnologías: el campo de la lingüística computacional". En: Fernández Pérez, Milagros (coord.), *Avances en lingüística aplicada*. Santiago de Compostela: Universidade de Santiago de Compostela, 147-227.
Ooi, Vincent (1998), *Computer Corpus Lexicography*. Edimburgo: Edinburgh University Press.
Payrató, Lluís; Boix, Emili; Lloret, M. Rosa e Lorente, Mercè (eds.) (1996), *Corpus, corpora*. Barcelona: PPU.
Pereira, Fernando e Warren, David (1980), "Definite Clause Grammars for Language Analysis". *Artificial Intelligence*, 13, 231-278.
Pérez Guerra, Javier (1998), *Análisis computarizado de textos: una introducción a TACT*. Vigo: Universidade de Vigo (Servicio de Publicacións).
Pérez Hernández, Chantal; Moreno Ortiz, Antonio e Faber, Pamela (1999), "Lexicografía computacional y lexicografía de corpus". En: Gómez Guinovart, Javier et al. (eds.), 175-213.
Pollard, Carl e Sag, Ivan (1994), *Head-Driven Phrase Structure Grammar*. Stanford: CSLI.
Ramallo, Fernando F. (1999), "Informática y sociolingüística cuantitativa". En: Gómez Guinovart, Javier et al. (eds.), 263-290.
Reiter, Ehud e Dale, Robert (1997), "Building Applied Natural Language Generation Systems". *Natural Language Engineering*, 3, 57-87.
Rosner, Michael e Johnson, Roderick (eds.) (1992), *Computational Linguistics and Formal Semantics*. Cambridge: Cambridge University Press.
Ruiz Antón, Juan Carlos (1996), "Modelos de análisis sintáctico en el procesamiento del lenguaje natural". En: Gómez Guinovart, Javier e Lorenzo Suárez, Anxo M. (eds.), *Lingüística e informática*. Santiago de Compostela: Tórculo, 31-55.
Shieber, Stuart (1988), "Separating linguistic analyses from linguistic theories". En: Reyle, Uwe e Rohrer, Christian (eds.), *Natural Language Parsing and Linguistic Theories*. Dordrecht: Kluwer, 33-68.
Solias, Teresa (1996), *Gramática categorial: modelos y aplicaciones*. Madrid: Síntesis.
Sperberg-McQueen, Michael e Burnard, Lou (eds.) (1994), *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Chicago: ACL-ACH-ALLC.
Strzalkowski, Tomek (ed.) (1999), *Natural Language Information Retrieval*. Dordrecht: Kluwer.
Tapias Merino, Daniel (1999), "Sistemas de reconocimiento de voz en las telecomunicaciones". En: Gómez Guinovart, Javier et al. (eds.), 83-102.
Vidal Villalba, Jesús e Busquets Rigat, Joan (1996), "Lingüística computacional". En: Martín Vide, Carlos (ed.), *Elementos de lingüística*. Barcelona: Octaedro, 393-446.
Wanner, Leo (ed.) (1996), *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam: John Benjamins.