Lexicalizaciones y expresiones idiomáticas. Su procesamiento

M.Meya
Siemens SA CDS
Barcelona

Resúmen

Los modismos y las frases idiomáticas son todavía un cajón de sastre poco sistematizado dentro de las teorías lingüísticas. Los giros, expresiones lingüísticas fijas son lugares comúnes en todo texto, y su significado no se puede calcular con las estrategias estandard de composicionalidad.

La propuesta que se presenta en este artículo se basa en que la estructura sintáctica de estas expresiones sigue los cánones de la sintaxis de una lengua dada. Los fenómenos de "distorsión" o "expansión" del significado se dan en función de la semántica de los elementos que constituyen dichas expresiones.

El significado de "llover a cántaros", "perder la cabeza", "sacar de quicio" , "a tontas y a locas" o "uña y carne", etc... se calcula primero a partir de la asignación de categoría, valencia o estructura determinada y luego contextualizando los elementos en juego.

Se distingue entre expresiones hechas verbales y no verbales. Las primeras se procesan siguiendo un reconocimiento de entornos o valencias, y se activan determinados procesos según se trate de un verbo funcional o de un verbo pleno.

Entre las expresiones no verbales se distinguen las tipo "fórmula" y las productivas que admiten la combinatoria con otros elementos entre medio. Las expresiones "fórmula" tienen entrada en el diccionario, mientras que las otras obtienen su significado a partir de las restriccioens semánticas que se dan en la expresión.

El valor semántico de una expresión es el valor semántico del núcleo más las anotaciones que aporta el contexto.

En un sistema de Traducción Automática o en una Interfaz en Lenguaje Natural no hay necesidad de diccionarios que contengan las entradas lexicalizadas; sin embargo, se necesita en el primer caso un diccionario de "transfer" que trasvase los significados de los elementos contextualizados a la lengua de salida, y en las interfaces se necesita una tabla-diccionario que convierta la representación primera a la deseada por la BD.

En definitiva, la propuesta presentada se apoya en las estructuras canónicas sintácticas de una lengua, y anota la información del contexto en los "elementos idiomatizados".

Lexicalizaciones y expresiones idiomaticas. Su procesamiento.

M.Meya

SIEMENS SA CDS Barcelona

.....

1 Introducción

Los modismos y las frases idiomáticas son todavía un cajón de sastre poco sistematizado dentro de las teorías lingüísticas. No obstante, para la **Traducción Automática** y en general para el **Procesamiento del Lenguaje Natural** su tratamiento se hace inevitable, ya que estas construcciones son muy frecuentes en la lengua y además son parte esencial de ella.

El problema fundamental radica no tanto en la problemática de como procesar estos "elementos complejos", sino en su identificación sistemática. Lo que para unos sistemas es una " lexía", o expresión idiomática y por tanto susceptible a ser entrada en el diccionario como una entrada única, para otros sistemas, o lingüistas la misma expresión se contempla como composicional.

El objetivo de este trabajo es proponer material de test para discernir una expresión idiomática de otra que no lo sea, crear por tanto, una tipología de los casos, y proponer un modo de procesamiento que posibilite el estabier equivalencias entre lenguas naturales con diferentes sistemas de lexicalización (lenguas románicas vs. lenguas germánicas).

El sistema que se propone para el procesamiento de expresiones complejas e idiomaticas, sacándolas de la esfera de la morfología y del diccionario,se decide en favor de un proceso que obtiene - en la medida en que eso es posible- el significado derivacionalmente a partir de los elementos que consituyen una unidad. La ventaja de una propuesta así, es que se consigue unicidad en el proceso, y a la vez se reduce enormemente la cantidad de entradas del diccionario.

Un sistema de TA multilingüe con un único diccionario de analisis debe contener sólo "unidades base", que en determinadas situaciones contextuales se traducan por un valor u otro en lengua destino. Si no se hace así se corre el riesgo de que los diccionarios sean inconsistentes, o que acaben siendo diccionarios fraseológicos sin motivación lingüística u orientados a los pares de lenguas en juego.

2 Lexicalizaciones productivas vs. lexías

El punto de vista que adopto en esta propuesta es el de la composicionalidad " When an expression admits analysis as morphologically or syntactically complex, assume as an operating hypothesis that the sense of the expression arises from the compostion of the senses of its parts". (Wasow, Sag,Nunberg 1982).

La consecuencia sintáctica de esta afirmación es que las expresiones idiomáticas no son diferentes de las "formas normales".

Las lenguas que tienen una determinada flexibilidad en el orden de las palabras continuan presentando en las frases idiomáticas estrechos lazos con las clases sintacticas a las que pertenecen. Una vez admitimos esto se llega a la conclusión de que las expresiones complejas e incluso las idiomáticas pueden ser tratadas con los mecanismos normales para el procesamiento del lenguaje.

Los principios de los que se parte para el procesamiento de expresiones complejas son:

- la descripción y comprensión de una frase idiomática -idiom- basada en un analizador sintactico normal - word driven-.
- la información idiomática que lleva a la distinción de significado frente a otra está introducida en las especificaciones del diccionario.

Con estos principios base en mente llegamos a dos tipos de construcciones: aquellas cuyo significado es la suma o el resultado de la interaccion de las palabras que constituyen esa unidad, y las expresiones o palabras compuestas cuyo significado es un valor añadido, y que no puede calcularse a partir de los componentes en juego.

Las primeras, las expresiones lexicalizadas, o lexías no pueden ser definidas en términos de sus componentes, mientras que las segundas sí. Aquellas tienen también unas características formales como:

son morfologicamente invariables, o son compuestos que no superan los test que avanzamos más adelante.

no admiten la inserción o cuñas de otros elementos entre sus constituyentes.

Una expresión que cumpla estas condiciones es una lexía y debe ser entrada en el diccionario con un asiento. En los diccionarios bilingües estas entradas tendrán correspondencias también lexicalizadas.

Las frases idiomáticas flexibles, por el contrario, pueden ser calculadas a partir de sus constituyentes, por lo que cada palabra es "desambiguada" en funcion del contexto en el que aparece.P.ej.: "número" en " números rojos" o entre " un trapo que sirve de estropajo" y " una persona sirve **de estropajo"** o la diferencia entre "planta" (floor) y "planta baja" (= ground floor)

La polémica reside en decidir si " planta baja" está lexicalizado o si el valor de "baja" se puede extraer del contexto, o si igualmente "estropajo" is a "mop" or a "worthless thing".

Siempre y cuando el significado de una unidad se mantenga en su esencia tanto si es como especificador o como especificado, entonces se dice que el significado se puede calcular composicionalmente. Los tests a realizar para llegar a una discriminación unívoca son:

Test 1 : El contenido de una unidad compleja dentro de una expre- sión se reformula sólo con el especificado **sin** el especificador. Si la segund formulación es aceptable, entonces el significado es composicional y **no** se le debe dar entrada en el diccionario.

P.ej.: Una planta baja es "una planta", mientras que no es así en

**Numeros rojos son números

Variantes de este test serían las propuestas por B. Jauregui. Si X es el especificado e Y el especificador y W el resultdo:

1. X e Y pertenecen a la misma clase

2. X es Y y viceversa

3. W es X

4. X es un medio, intrumento, causa, contenedor o resultado de Y

Ejemplos:

1. Un vaso probeta ; una mujer bruja

2. autor-editor : pianista compositor

3. reloj de bolsillo, estación de autobuses, planta acuática,...

4. cuchara de sopa, reloj de bolsillo

Como se pude observar los ejemplos bajo los apartados 1-4 son productivos y podremos crear multitud de ellos fácilmente. Se trata de construcciones productivas y sería absurdo incluirlas en el diccionario.

No obstante, podemos considerar ejemplos similares en donde lexicografos estarían en duda. Esta situación es todavía más plausible si comparamos entradas en diccionarios multilingües:

"Atomabstand" - interatomic distance - distancia interatómica

"Atommüll" - radioctive waste - residuos radioctivos

"Atomrumpf" - atomic residue - tronco del átomo

"Atomwaffen" - nuclear weapons - armas nucleares

Si el significado de **"Atom"** oscila entre : **interatómico, radiactivo, átomo, y nuclear** ¿ en qué medida esto es composicional, y en qué medida está lexicalizado con la unidad a la que se une ? Menos en el caso de "Rumpf" el test 1 funcionaría

¿ Entonces cuándo significa el especificador **"Atom"** en los compuestos de los anteriores ejemplos "radioactico" y cuándo "nuclear" ?

Si adoptamos el punlo de vista monolingüe nunca se decidiría un lexicógrafo a incluir como entrada del español :"residuo_radioactivo", "carga_radioactiva", "arma_nuclear",...

Si es así, una vez se ha decidido que "armas nucleares" son un tipo determinado de armas, y que **"llover a cántaros"** es una manera de llover, lo mismo que "servir de estropajo' es una manera de ser utilizado, pero que **"estar en números rojos"** no es estar en determinados números, y que **"sacar de quicio"** no tiene nada que ver con el "quicio de una puerta"..., entonces estamos en condiciones de separar las entradas **lexicalizadas** de aquellas que **no lo son.**

El siguiente paso se trata de caracterizar los tipos y encontrar un modo de procesamiento siguiendo los cánones del análisis sin- táctico.

3 Tipología de las expresiones complejas composicionales

La mayoría de las expresiones composionales son productivas en el sentido en que en la lengua se pueden crear espontáneamente a partir de modelos establecidos. He aquí unos ejemplos de los casos frecuentes:

CONSTRUCCIONES CON PARTICIPIO

orientado a ....."dialogorientiert:

distribuido a ...

protegido_por "Metall**geschuetzt**"

CONSTRUCCIONES ADVERBIALES CON PPs

en_cascada ... Kaskadier**bar**

en_forma_de ... **blockweise**

en_serie ... bitseriell, **hintereinanderschaltbar**

a_prueba_de ... krichstrom**fest, waterproof**

sin_... ... Koffein**frei**

según_... ... arbeits**maessig**

CONSTRUCCIONES ADVERBIALES CON ADJETIVOS

intensivo_en ...arbeits**intensiv**

fácil_de ...einfach zu handhaben

dependiente_de ...richtungs**abhaengig**

bajo_en ...Nikotin**arm, (low)**

Estas construcciones se expresan en las lenguas germánicas o bien mediante sufijos, o bien con ADJ que en las lenguas románicas se expresan mediante grupos preposicionales.

CONSTRUCCIONES CONJUNTIVAS/DISYUNTIVAS AGLUTINADAS

Este tipo de construcciones se dan sobre todo en los textos técnicos o con carácter telegráfico.P.ej.:

Datos de organización/administrción

Datos de entrada-salida

CONSTRUCCIONES CON ACRONIMOS ESPECIFICADORES

micro_b_display ...indicador_micro_B

master slave ...maestro-esclavo

APOSICIONES

código_MICR ...MCR code

circuito flip-flop

configuración maestro-esclavo

CONSTRUCCIONES CON GRUPOS NOMINALES

MIC IN ..entrada de micrófono

master ON OFF ..conmutador principal

CONSTRUCCIONES IDIOMATICAS VERBALES

Las expresiones idiomaticas construidas en torno a un verbo se subdividen en aquellas que se crean con un verbo funcional, verbo gramatical quasi-vacio, y las que se forman con verbos plenos.

verbos funcionales son: hacer, ser, estar, tener, dar, ofrecer, efectuar, realizar,etc...

Las construciones con los verbos funcionales da una triple tipología de construcciones complejas o idiomaticas:

(1) el VRB y el objeto directo forman una unidad de significado

Se trata de los lexemas verbales complejos propios del lenguaje periodístico.

efectuar : el pago, el requerimiento, la concesión, un disparo,...

hacer : alusión a, aparición, una exposición, denuncia,..

(2) la expresión resultante no es paráfrasis del verbo:

Los NP/PP mantienen su significado primitivo.

ponerse en movimiento

estar en apuros

(3) El significado standard del verbo funcional o pleno se mantiene pero es el NP el que lleva la carga del significado. Esto obliga en un sistema de traducción a generar un substantivo completamente distinto. Se trata de expresiones metafóricas

tener resaca ....ein Kater haben

estar en numeros rojos ...in der Kreide stehen
dar en el clavo ...ins Schwarze treffen

estar a las últimas ...

estirar la pata ...verrecken

practicar la política de avestruz

<u>En resumen:</u>

Grupos nominales: el significado se calcula a partir de las especificciones que aportan los PPs, AdJs y ADVs.

Grupos verbales : con verbos gramaticales, el signifcado se distribuye en 4 casos distintos según esté la carga semántica en los complementos del verbo o en este, o sea algo independiente, no composicional.

4 Tipología de las expresiones idiomáticas

Una expresión idiomática tanto nominal como verbal es aquella cuyo significado no se calcula composicionalmente.

Como en el caso anterior las construcciones se subdividen en:

Nominales : (no-verbales) estas pueden ser de : **Tipo fórmula, nominales complejos y compuestos**

Verbales : expresiones metafóricas de significado acompo- sicional tanto con verbos plenos como funcionales.

4.1 Expresiones idiomaticas nominales

La contrapartida de los ejemplos productivos vistos antes está en ejemplos técnicos cuyo significado es ciertamente la función de las palabras sobre la que se aplica.

NPs complejos : (ejemplos)

Derivación con enlace directo (metal-to-metal-tap)

cable-del-contador-de-impulsos (meter wire)

recepción de emisión en banda reducida (microlock)

Una expresión es de naturaleza idiomática siempre y cuando sea imposible insertar una palabra entre las unidades que la componen.

Este es siempre el caso en las llamadas "fórmulas" léxicas. Estas son **unidades indivisibles** sintácticamente.

a) pares

blanco y negro

carne_y_uña

dueño_y_señor

b) formulas comparativas

como sardinas en lata

como pez en el agua

c) lugares comunes en sublenguajes y expresiones adverbiales

de_mal-en_peor

lamentándolo_mucho

a_fin_de_cuentas

referente_a

Lo característico de estas unidades es su indivisibilidad sintáctica. Su procesamiento ,obviamente, las considera como una unidad.

4.2 Modismos verbales

Estas expresiones son las más cuantiosas, y todas las lenguas son ricas en ellas. Tanto las que se construyen con verbos funcionales como hemos visto más arriba, como las que se crean con verbos plenos, todas siguen las restricciones sintácticas correspondien- tes al verbo. Es decir, si un verbo tiene asociado dos o tres entornos sintacticos, las construcciones figurativas o idiomáticas quedan encuadradas en el mismo entorno.

Ej.: "estar" ...tiene un entorno prototípico que es el LOCal.

Juan está en Barcelona // expr. normal

Juan está en babia // expr. idiomática

"chupar" ... exige OBJeto directo. la niña se chupa la herida // expr. normal

la niña no se chupa los dedos // expr. idiomática

"perder" ... exige también un OBJeto directo

perdió la paciencia // expr. normal

perdió los estribos // expr. idiomática

A nivel lingüístico los modelos sintácticos que aparecen en las expresiones idiomáticas son los prototípicos del verbo. La única variante son las desviaciones en los valores semánticos de los complementos y suplementos que toman las posiciones de las valencias.

Ej.: "llover" admite adverbiales de modo

"llueve en cantidad" vs. ""llueve a cántaros"

La ventaja de definir estos valores metafóricos en función de los prototípicos del verbo es que no necesitan entrada en el diccionario monolingüe de un sistema de traducción.

La propuesta que se presenta definiría al verbo "llover" con sus entornos normales. "a_cantaros" sería un adverbio normal, una unidad indivisible. El analizador sintáctico cumpliría la tarea de reunir en un constructo sintáctico a ambas unidades.

5 El diccionario

El diccionario de un sistema para el procesamiento del lenguaje natural contendría únicamente las unidades múltiples que estuvieran lexicalizadas. Es decir, aquellas formas simples o los compuestos que fueran el significante de un concepto. La estrategia para dar razón de la combinatoria de estas unidades tanto estandard, como idiomáticas sería tarea del analizador sintáctico y de las restricciones semánticas añadidas.

Algunos sistemas para la TA, o para la comprensión del lenguaje Natural tienen entradas complejas aun cuando estas son susceptibles a admitir otras palabras entre medio. Esto hace que sus diccionarios sean poco transparentes, es decir, difíciles de manejar y ampliar, y además al ir aumentando sus aplicaciones los asientos que dan razón de un **mismo concepto** van creciendo exponencialemnte, ya que deben incluir todas sus variantes contextuales.

P.ej.: "Ayer llovió a cántaros"

"Ayer llovió en Barcelona a cántaros"

"En Barcelona ultimamente llovió a cántaros", etc

Con estos tres ejemplos tan simples podemos ver lo fácil que es encontrar cuñas léxicas entre lo que en el diccionario se ha definido como una **unidad idiomática** que se reonoce como una **cadena continua**. Al ser así, el diccionario irá creciendo al tener que incluirse entre 'llover' y' a cántaros' todas las variaciones del tipo que hemos visto en los ejemplos de más arriba.

Una estrategia así, es decir de **pattern-matching** y no lingüísica es la que siguen sistemas como ALPS, WEIDNER y SYSTRAN. Un sistema, que por el contrario, esté basado en un conocimiento lingüístico más sofisticado habrá desarrollado modelos más complejos para la fase de reconocimeinto, pero las estructuras serán 'repetibles', y tendrá la enorme ventaja de reducir el diccionario en proporciones incomparables.

Un diccionario como el que se propone aquí es el que se usa en el sistema de TA METAL. Tomando como punto de partida los ejemplos que hemos presentado en este artículo, el diccionario tendría entradas para:

llover VRB frame_1 - frame_n

estar VRB frame_1 - frame_n

en_cantidad ADV

a_cántaros ADV

número NOM

rojo ADJ frame_1

carne_y_uña NOM

etc...

No habrían, sin embargo, entradas complejas (multi-words) para verbos con lec-turas metafóricas o con nominalizaciones (tipo:efectuar una detención = detener), ni para grupos nominales complejos que recogen el significado de un concepto. Cuando es asi, las traducciones de estas unidades a otra lengua corresponden tanto a compuestos como a simplex.

logische Verbindung ... sesión

watermill ... molino de agua

sailboat ... barco de vela

garden city ... ciudad jardín

automatic closed-loop control...

5 El procesamiento

El procesamiento de las expresiones idiomáticas dentro del modelo propuesto sigue estrictamente las directrices del análisis de las construcciones estandard. La información requerida está en el caso de las construcciones verbales en el verbo. Hay un diccioanrio que da la información acerca del tipo de entorno sintáctico que requiere el verbo; otro diccionario nos da el valor semántico de la construcción en función de la base de datos que tengamos. Se trata de una traducción a otro nivel. Esta traducción se hace a partir de la información adquirida a largo del proceso de reconocimiento. Para cada valor adquirido hay una sustitución que nos da el valor en la lengua de salida correspondiente, o en el mundo de la BD en que nos movamos.

En las lenguas germánicas los compuestos no lexicalizados, es decir aquellos en que el análisis los descompone en sus elementos, son procesados, como toda expresión compleja en una lengua románi- ca marcando su función y añadiendo a su información estructural el tipo semántico del especificador en el nucleo, y la del nucleo en el especificador. Se trata de saber, que especificadores acompañan a un núcleo nominal y de qué tipo son. He aquí dos ejemplos paralelos uno para compuestos alemanes, ingleses y sus correspondientes en español.

ej.:1a. "logische Verbindung" .....

1b " sesión"

2a. "eheliche Verbindung" .....

2b " relación matrimonial"

3a. "Telephonverbindung" .....

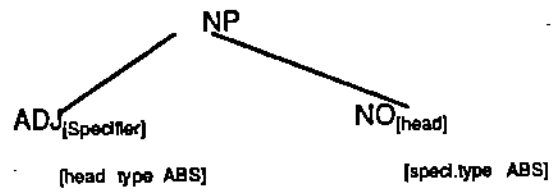
3b " conexión telefónica"

4a. "Pressluftlhammer" .....

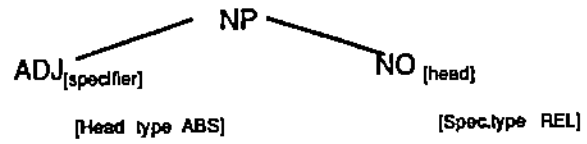
4b " martillo hidráulico"

Los esquemas siguientes recogen las categorias sintacticas que han sido anotadas a lo largo del analisis.

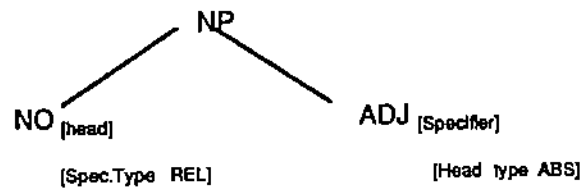
1a.



2a



2b



Durante la fase de análisis el analizador aparte de reconocer los componentes de cada constituyente sintáctico copia los rasgos se-mánticos del núcleo al especificador, y los del especificador al núcleo. El elemento en cuestion **'anotado'** (decoration) con estos valores es obviamente en cada caso una entidad semántica diferente.

En la fase siguiente, los diccionarios de asignaciones, o en su caso los procesos composicionales obtienen los valores de salida correspondiente. En el caso de un diccionario bilingüe, a cada uno de los valores anotados le corresponde una entidad léxica,(también anotada) diferente.

Así, la palabra "Verbindung" se traduce por **'relación'** o **'conexión'** según sean las anotaciones obtenidas del análisis del contexto. El proceso inverso, de pasar de una expresión múltiple en castellano a otra múltiple, a una forma aglutinada o a un compues-to en una lengua germánica sigue los mismo principios.

En los casos de construcciones verbales idiomáticas son las restricciones semánticas de los entornos verbales realizados en la frase/oración las que deciden acerca del significado final de la expresión. Si se trata de un verbo funcional, se 'anotan' con su valor sus argumentos/complementos. Estas anotaciones capacitan al sistema a obtener el significado final.

En el ejemplo "Der Mann bringt mir auf die Palme"

"El hombre me saca de quicio"

Al descubrir el sistema que **"bringen"** es un verbo funcional, entonces marca a sus constituyentes adecuadamente. Al traducir **"Palme"** (palmera) el sistema sabe que apareció en un contexto en que estaba con el verbo 'bringen'.

Un sistema que tuviera que traducir al alemán, tendría en ese momento contextualizados sus elementos, y estaría en condiciones de hacer el 'transfer' o traducción a la lengua destino. si se tradujera al alemán "Palme" por aparecer en una frase idiomatica no se traduciria por "palmera" sino por "quicio".Esto ocurre porque "Palme" esta contextualizado por "bringen". He aquí, que de esta manera tan simple habríamos hecho un trasvase de información idiomática de una lengua a otra.

La grafica 1 muestra la estructura del analisis reconocido por el sistema de Traduccion Automatica METAL. Dentro de los rasgos que tiene

el nodo NP "Palme"(palmera) existe el de hacer referncia al verbo funcional "bringen" con quien se realiza en la oracion. Debido a estas "anotaciones" (rasgos) el sistema puede ir transformando pieza a pieza las construcciones idiomaticas.

RESUMEN

La propuesta presentada en este artículo se apoya en tres pilares:

- tratar las expresiones idiomáticas en la medida posible con los mismos criterios que las construcciones gramáticales estandard.
- localizar la función de cada uno de los elementos implicados

Con ello se obtienen dos (a),(b) ventajas:

- a) las BD léxicas son independientes de la aplicación
- b) el trasvase '**Idiomático**' queda reducido a la aplica- ción dada (a los pares de lenguas en juego si se trata de traducción automática)
- - Copiar los valores semanticos del especificador en el núcleo y los del núcleo en el especificador para así poder aplicar restricciones seleccionales.
- - Contemplar la idiomatización como algo relacionado entre dos códigos, y no por su naturlaeza intrínseca. Es decir, algo es idiomático si en otro sistema sigue otros códigos;

Si los códigos son similares o idénticos una expresión quedará en ser '**compleja**'(consta de múltiples elementos) pero se podría tratar composicionalmente en función de la aplicación en juego.

Si dos expresiones se construyen de forma paralela en dos lenguas como es el caso para los ejemplos: "construir castillos en el aire" o en alemán: 'Luftschloesser bauen' la traducción se limitará a traducir uno a uno sus elementos.

Si por el contrario los códigos de expresiones de este tipo son distintos esta expresión a parte de ser compleja sería 'idiomática'. Este es el caso de la metáfora correspondiente en frances, a otras unidades léxicas. Por ejemplo, la traducción que se traduce por:' bâtir chateaux en Espagne'.

****

BIBLIOGRAFIA

A. Schenk : *Idioms in the Rosetta Machine Translation System.* En. Proceedings of the COLING' 86. Bonn p.319

U. Zernik :*Disambiguation and Language Acquisition through the Phrasal Lexicon.* Proceedings of COLING'86. Bonn p. 247-252

J.Howard Shaw: *Motivierte Komposita in der deutschen und englischen Gegenwartsprache.* G.Narr.Tuebingen 1979