Geographic reference analysis for geographic document querying

Frédérik Bilhaut, Thierry Charnois, Patrice Enjalbert and **Yann Mathet**
GREYC, CNRS UMR 6072, Université de Caen
Campus 2, F-14032 Caen Cedex – FRANCE
{fbilhaut,charnois,patrice,mathet}@info.unicaen.fr

Abstract

The work presented in this paper concerns Information Retrieval from geographical documents, i.e. documents with a major geographic component. The final aim, in response to an informational query of the user, is to return a ranked list of relevant passages in selected documents, allowing text browsing within them. We consider in this paper the spatial component of the texts and the queries. The idea is to perform an off-line linguistic analysis of the document, extracting spatial expressions (i.e. expressions denoting geographical localisations). The point is that such expressions are (in general) much more complex than simple place names. We present a linguistic analyser which recognises them, performing a semantic analysis and computing symbolic representations of their "content". These representations, stored in the text thanks to XML annotation, will act as indexes of passages with which queries are compared. The matching of queries with text expressions is a complex process, needing several kinds of numeric and symbolic computations. A prospective outline of it is described.

1 Presentation of the GeoSem project. Passage extraction from geographical document

The work presented in this paper concerns Information Retrieval (IR) from geographical documents, i.e. documents with a major geographic component. Let's precise at once that we are mainly interested in human geography, where the phenomena under consideration are of social or economic nature. Such documents are massively produced and consumed by academics as well as state organisations, marketing services of private companies and so on. The final aim is, in response to an informational query of the user, to return not only a set of documents (taken as wholes) from the available collection of documents, but also a list of relevant *passages* allowing text browsing within them.

Geographical information is spatialised information, information so to speak anchored in a geographical space. This characteristic is immediately visible on geographical documents, which describe how some phenomena (often quantified, either in a numeric or qualitative manner) are related with a spatial and also, often, temporal localisation. Figure 1 gives an example of this informational structure, extracted from our favourite corpus (Hérin, 1994), relative to the educational system in France. As a consequence a natural way to query documents will be through a 3-dimensional topic, Phenomenon-Space-Time as shown in Figure 2. The goal is to select passages that fulfil the whole bunch of criteria and to return them to the user in relevance order.

The system we designed and currently develop for that purpose is divided in two tasks: an off-line one, devoted to linguistic analysis of the text, and an online one concerning querying itself. Let's give an overall view of the process, focusing on the spatial dimension of texts and analysis. Other aspects of the project, including especially the analysis of expressions denoting phenomena, techniques used to link the three components of information (Space, Time, Phenomena) and implementation issues can be found in (Bilhaut, 2003).

Concerning text analysis, the goal is to locate, extract and analyse the expressions which refer to some geographical localisation [1] so that they act as *indexes* of text passages. The first remark to do is that we have to cope (in general) with complex nominal expressions, not only named geographical entities, as exemplified in figure 3. Indeed the collection of (proper) place names can

[1]Temporal expressions (expressing temporal localisation) are treated in a similar manner.

De 1965 à 1985, **le nombre de lycéens** a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible **dans le Sud-Ouest et le Massif Central**, modérée **en Bretagne et à Paris**, l'augmentation a été considérable **dans le Centre-Ouest, et en Alsace**. [...] Intervient aussi **l'allongement des scolarités**, qui a été plus marqué dans les départements où, **au milieu des années 1960**, **la poursuite des études après l'école primaire** était loin d'être la règle.

***From 1965 to 1985**, **the number of high-school students** has increased by 70%, but at different rythms and intensities depending on academies and departments. Lower in **South-West and Massif Central**, moderate in **Brittany and Paris**, the rise has been considerable in **Mid-West and Alsace**. [...] Also occurs **the schooling duration increase** which was more important in departments where, **in the middle of the 60's**, **study continuation after primary school** was far from being systematic.*

Figure 1: Excerpt from (Hérin, 1994)

Find the passages which concern:

- Le retard scolaire dans l'Ouest de la France depuis les années 1950.
- Educational difficulties in West of France since the 50's.

- L'évolution des effectifs dans l'enseignement secondaire à Paris / dans la région parisienne.
- Variations of the number of pupils in secondary school in Paris / in Paris area

- L'évolution des effectifs scolaires dans les régions rurales.
- Variations of the number of pupils in rural areas.

- Les mutations du personnel enseignant dans les académies du Sud.
- Transfers of the teaching staff to southern districts.

Figure 2: Typical queries on geographical documents.

not constitute an adequate index: a mention of "north of Paris" or "north of France" has obviously not the same meaning as "Paris"or "France", not to speak of "south of a Bordeaux-Genève line". Moreover, some expressions ("industrial towns" or "rural departments"...)[2]" involve a "qualitative" (demographic, sociological, economic...) characterisation of the selected areas, involving some knowledge of this kind.

The conclusion is that a literal matching of "queries" against "text expressions" simply can't do. Expressions (and queries) must receive a linguistic analysis, discovering their structure and producing some kind of *semantic representation*. This is the goal of the off-line text processing step. A linguistic analyser of spatial expressions (nominal and prepositional phrases) have been designed, which recognise them and produces a symbolic representation of their "content". These representations are associated with the text, thanks to XML annotation, and constitute the index with which queries will be compared. The linguistic analysis is described in section 2.

Assuming that such an analysis is performed, we are ready for querying. Clearly the easier way for a user to formulate his/her query is to use also natural language. The first step will be to apply the same linguistic analysis, producing a symbolic representation of the same nature as what was extracted from text. We have then to perform some matching between (the representations of) the query and the text. This is not a trivial task, as the reader can guess, considering expressions and queries in figures 2 and 3. To achieve this task, we will use referential information associated with named geographical entities (long-lat coordinates) together with some computation exploiting the symbolic representations produced by the linguistic analysis. A (prospective) sketch of this process is described in section 3.

Summing up to situate the project among current research, we see that the goals are those of Document Retrieval, but at an intra-document level, selecting passages (Callan, 1994). But the methods are rather (though not exclusively) those of Information Extraction in the sense of MUC's (Pazienza, 1997) and we are quite close to Answer Extraction in the sense of (Molla, 2000). In particular, the spatial component of geographical texts needs much more than an access to geographical resources as gazetteers: it needs both a specific semantic analysis of complex linguistic expressions, and some symbolic and numeric spatial computation for matching the query with text. Let's now consider these two aspects in turn.

[2]"departments" denotes in France administrative districts, roughly equivalent to "counties"

QUANT	TYPE		ZONE	
	administrative	*qualification*	*position*	*named geo. entity*
(1)				à Paris
(2)			au nord de	la France
(3) Quelques	villes	maritimes		
(4a) Le quart des				
(4b) Tous les	départements		du nord de	la France
(4c) Quelques				
(4d) Quinze				
(5) Quelques	villes	maritimes		de la Normandie
(6) Les	départements	les plus ruraux	situés au sud de	la Loire

(1): in Paris
(2): in north of France
(3): some seaboard towns
(4a/b/c/d): The quarter of / All / Some / Fifteen / districts of north of France
(5) Some Seaboard towns of Normandy
(6) The most rural districts situated from south of Loire

Table 1: Structure of spatial expressions

- Paris

- Les villes industrielles d'Île de France.
- *Industrial towns in Île de France.*

- La moitié nord de la France.
- *The northern half of France.*

- Les départements ruraux du nord de la France.
- *Rural departments in the north of France.*

- Au sud d'une ligne Bordeaux-Genève.
- *In the south of a Bordeaux-Genève line.*

Figure 3: Typical spatial expressions

2 Spatial analysis

2.1 Description of spatial expressions

Table 1 shows a significant sample of spatial expressions found in our corpus. It enlightens the two components which characterises their informational structure, [Type] and [Zone], Which can be altogether present or not. Hence three kinds of expressions can be considered:

1. Expressions in the first class contains only the [Zone] part and denotes a georeferenced area (examples (1) and (2)). They are anchored in a named place (Caen, France, Normandy...), later called "named geographical entity" (egn), on top of which some spatial, geometrical, operations can act (north/south of, the western/eastern part of, the surroundings of...).

2. The second type of expressions denote a set of places and can be summarised by the canonical form [QUANTIFICATION]+[TYPE]+[ZONE]. The set is quantified by a determiner (all, some, the, most of...). The places are generally given an administrative type (town, region...), and are located in a zone. Sometimes, a further qualification, either sociological (rural, urbanised, more or less densely populated, ...) or physical (near seaboard, mountainous, ...) specifies the type. We call this most general type of expressions "LocGeo" (geographical localisations): examples (4a–d),(5),(6).

3. Finally, the form [QUANTIFICATION] + [TYPE] is a variant of the second form with a zone not expressed but implicit (and dependent on the context). Note that this kind of expression is recognised by our analyser if the qualification field is present. That means that expressions as "the districts" are not considered as a geographical entity in opposite of "the most rural districts" which can be geographically determined: example (3).

Note that all components of the expression must be taken into account when the semantic representation is computed: not only the elements of geographical type but also the quantification, the qualification, the type which are crucial for querying and matching the representation (see section 3).

2.2 Semantic representations

The semantics of expressions is represented by feature structures as shown in figure 4.

$$(1)\quad \begin{bmatrix} zone : \begin{bmatrix} egn : \begin{bmatrix} ty\_zone : ville \\ nom : Paris \end{bmatrix} \\ loc : interne \\ coord : \begin{bmatrix} lat : 45.6333333 \\ long : 5.7333333 \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Paris

$$(4b)\quad \begin{bmatrix} quant : \begin{bmatrix} type : exhaustif \end{bmatrix} \\ type : \begin{bmatrix} ty\_zone : departement \end{bmatrix} \\ zone : \begin{bmatrix} egn : \begin{bmatrix} ty\_zone : pays \\ nom : France \end{bmatrix} \\ loc : interne \\ position : nord \end{bmatrix} \end{bmatrix}$$

Tous les départements du nord de la France

$$(5)\quad \begin{bmatrix} quant : \begin{bmatrix} type : relatif \end{bmatrix} \\ type : \begin{bmatrix} ty\_zone : ville \\ geo : maritime \end{bmatrix} \\ zone : \begin{bmatrix} egn : \begin{bmatrix} ty\_zone : region \\ nom : Normandie \end{bmatrix} \\ loc : interne \end{bmatrix} \end{bmatrix}$$

Quelques villes maritimes de la Normandie

Figure 4: Spatial expressions accompanied by their semantic representation

The “quant” feature corresponds to the quantification [QUANT] part of expressions, expressed by generalised determiners. It will be used to associate some approximate cardinality to the set of elements selected in the LocGeo, allowing to compute some “relevance value” with respect to a given query (see section 3). Four types of quantification are distinguished:

- absolute: “fifteen districts”
- relative: “32 per cent / half of the towns”
- exhaustive: “all”, “the”
- exhaustive-negative: “no”, “not any”

The “type” field gives the qualitative characterisation of the selected places. It can be administrative (ty_zone) and/or socio-economic (geo).

Finally the “zone” feature gives an abstract description of the geographical localisation. It is defined by four possible sub-features:

- “egn” (named geographical entity) with the name and type of this entity;
- the “coord” field gives the coordinates of the named place, when available.
- when expressed, the “position” describes the spatial operator acting on the egn (for example a N-S-E-W orientation);
- this information is completed by the “loc” feature (localisation), which can take only two values, internal or external, according to whether the geometrically selected area ”position” criterium applies inside or outside the zone: “in the north of France” is internal to France, while “ north of Paris” is (probably) external.

  It must be mentioned here that his last feature raises interesting and difficult semantical considerations and the implemented procedures yet subject to strong limitations. Notably, it is strictly local (it does not consider the context of the analysed expression), while the localisation can be ambiguous: “north of Paris” can also denote an internal (northern) part of Paris. More generally, a precise study of spatial prepositions remains to be done.

The semantics of extracted phrases (represented as feature structures) are exemplified in Fig. 4. Example (4.b) stipulates an exhaustive determination selecting all entities of the given TYPE (departments) located in ZONE. This zone matches with the northern half inside the named geographic entity (France). In (5) the determination (induced by "quelques / *some*") is relative, i.e. only a part of the elements given by the type is to be considered. Here, TYPE stipulates that we only keep from ZONE (Northern Normandy) the "towns" which are "seaboard".

In fact, the structure (and semantics) of spatial expressions is significantly more complex allowing notably:

- some kind of recursivity as in: "les villes maritimes des départements ruraux du nord de la France" (*the seaboard towns of rural districts in north of France*) where the LocGeo “villes maritimes” is embedded inside the LocGeo “départements ruraux du nord de la France”.
- Geometrically defined zones: "le triangle Avignon-Aix-Marseille" (*the Avignon-Aix-Marseille triangle*); and areas defined by some kind of bounds: "du Sud-Ouest à la Bourgogne" (*from South-West to Burgundy*).
- Enumerations of different kinds: "à Paris, Lyon et Marseille" (*in Paris, Lyon and Marseille*), "dans les départements de Bretagne et de Normandie" (*in the departments of Brittany and Normandy*), "dans la

France du Centre et de l'Ouest" (*in Center and West of France*), etc.
Such enumerations are quite frequent in the corpus, and treated by the linguistic analyser. In particular the different entity types appearing in an enumeration are considered, and simultaneously the lexical head of the phrase is correctly distributed over the constituents of its expansion, as in: "dans les départements bretons et normands, à Paris, et dans les régions du sud et du sud-ouest" where "départements" is distributed over "bretons" and "normands", and similarly "régions" over "sud" and "sud-ouest".

- some anaphoric expressions not still treated ("ces régions", *these regions*).

2.3 Implementation

The whole process is implemented using the Lingua Stream platform, designed for the project [3]. We assume a tokenisation and a morphological analysis of the text: presently we use Tree-Tagger (Schmid, 1994) which delivers the lemma and part-of-speech (POS) categorisation. This is turned into a form acceptable by Prolog (list of terms) and a definite clause grammar (DCG) performs altogether syntactic and semantic analyses. The semantic representations are synthesised in a compositional process. Prolog proves to be an interesting choice here since it allows complex semantic computations to be integrated in the grammar, and unification on feature structures thanks to GULP (Covington, 1994). Presently, the grammar contains 160 rules and an internal lexical base of about 200 entries, including grammatical words and administrative or socio-economic terms. A gazetteer of 10000 named places located in France is used as external lexicon, providing administrative types and geographical coordinates[4].

2.4 Evaluation and results

The analyser was designed by observation of (Hérin, 1994), and a qualitative evaluation on several other texts seems to indicate that we captured correctly the general structure of spatial expressions. However a more precise, quantitative, evaluation on a wide and diversified corpus is still an open question, left for further work. Another important aspect concerns evaluation of the semantic analysers, esp. the spatial one. We have to compare the semantic structures computed by the system with expected ones and hence to define a relevant and robust measure of adequation between complex feature structures.

[3]publicly available from the web site: http//users.info.unicaen.fr/ fbilhaut/linguastream.html

[4]publicly available from the web site: http://www.nima.mil/gns/html/

The whole process takes about 30' on our favourite corpus (Hérin, 1994) (200 text pages). 900 expressions are recognised. Though processing time is not so crucial for off-line analysis, we also want to improve the system's efficiency: working on the grammars and their implementation techniques (such as bottom-up parsing and compilation of feature structures as described in (Covington, 1994)),we hope to gain a factor 2 or 3. Other, possibly more efficient, parsing methods could also be considered if necessary, provided a good integration in the LinguaStream platform is preserved.

3 Semantic matching

The work presented now is mainly prospective. As announced in part I, once a user has entered a query, the system has to determine whether a given passage is relevant or not. More precisely, we expect the system to deliver, rather than a yes/no value, a relevance degree, so that a ranked list of passages (from best to least) would be returned to the user as an answer to his/her query. This task requires to perform some matching between the semantic representation of the request (calculated on-line) and the one of a given passage (precalculated, i.e. off-line). Figure 5 shows a query (Q1) and six passage excerpts (from P1 to P6) so that we have an overview of the problem : for each passage, is this passage relevant, how relevant is it, and according to what criteria ?

Q1) Which passages address Paris ?

P1) La capitale [...] - *the capital (city)*.
P2) Les villes de la Seine à l'exception de Paris - *Towns in the department of Seine, Paris excepted*
P3) Les grandes villes françaises- *Big cities in France*.
P4) La moitié nord de la France - *The northern half of France*.
P5) Au sud d'une ligne Bordeaux-Genève - *South of a Bordeaux-Genève line*.
P6) La plupart des villes d'Ile de France - *Most of towns in Ile de France*.

Figure 5: One query and six passage excerpts

We can split the process in two major steps. First, a compatibility diagnosis is delivered. In other words, it consists in stating whether a passage could be relevant or not, relying on the fact that there is no geographic incompatibility. This step should obviously ban P2 and P5. Second, if a given passage is considered as "compatible", the system computes a relevance degree, i.e. delivers a value from zero (worst, shouldn't happen if the first step does its job) to one (best, you can't find a more relevant passage with respect to the query). This step would deliver a ranked list such as [P1, P3, P6, P4]. (The precise order is still a open question, since it is not obvious that P3 should come before P6, for instance).

Let's go deeper in these two steps.

3.1 Compatibility computation

From examples in Figure 5. we can immedialtely see that this task involves in the general case much more than a simple word matching between the query and text expressions. Otherwise, the system would only be able to return passages which contain "Paris", namely P4, which precisely should be banned as result of the negation form ("Paris excepted"). Let's present what kind of knowledge and algorithms the system can use to process query Q1 relatively to passages P1 to P6.

P1: In the context this excerpt is taken from, we are concerned with France. A gazetteer or a GIS would clearly say that France's capital is Paris. Hence, the answer is quite easy to get but, once more, doesn't rely on direct word matching.

P2: This excerpt has a two components structure, namely "les villes d'Ile de France" and "à l'exception de Paris". The first one needs an answer to :

- a "is type of X T?" question, X being here Paris, and T "city", and
- a "does X belongs to Y ?" question, X still being Paris, and Y being a french department, la Seine.

For both these questions, a gazetteer will do (and answer positively). The second component "à l'exception de Paris" states a restriction over the first requirement and, since "Paris" matches "Paris", P2 would finally be banned.

P3: Paris is a french city, as a gazetteer would say. But we're concerned here with more sofisticated semantics as we have to interpret "big cities" (a piece of knowledge not in the scope of a gazetteer). Since qualifiers such as "big", "middle", "small" and so on, are relative to a set of entities, we propose to generate off-line a resource such as the rank of a X entity relatively to a criterium C among all entities of a given type T. We can then interpret "big" as "to be in the upper 20%", and so on with "medium" or "small". Note that french qualifier "grand" is quite ambiguous, (denoting population as well as surface), but "largest" clearly involves surface criterion.

In the next two examples, we'll have to go deeper in geographic computation. Indeed, we understand that P4 is compatible with Q1, and that P5 isn't, but how could the system know that ? It has to compute it. This requires a kind of geometric compatibility between shapes associated with the entities denoted by the request and text expressions, in topological terms: does X cover (or intersect) Y ? Contrary to examples P1 to P3, where this compatibility was *de facto*, embedded in gazetteer-like knowledge, we have now to cope with complex and dynamic geographic compatibility that no gazetteer nor GIS could directly deliver : "the northern half of France" nor "the south of a Bordeaux-Genève line" couldn't be indexed in a database.

P4: "northern half of X" requires to cut the area associated with X in two, so as to keep only the part situated above the middle line. Here, X is France. A request to a GIS gives the minimum and the maximum longitude, so that we obtain the middle line, and finally the shape as given in Figure 6. Then, we look at the long-latt coordinates of Paris, and conclude that Paris matches this passage.

P5: For similar reasons, but with a more sophisticated shape, we obtain a polygon as given in Figure 7. And we conclude here with a no answer, since the coordinates of Paris locate it out of the polygon.

Figure 6: The northern half of France

We now evoke briefly other problems which must be faced. First, observe that there are some subtilties in the compatibility relation, as illustrated in the following example. How to interpret fluzzy relations like "north of X" or "south of X" ? What part of the area associated with X do we have to keep ? Further work must be done to make the best choice between something like Figure 6 where we keep half of it, and another one where we would keep a smaller part, or for exemple a cone above X. How to interpret the french "nord de X" in terms of topological "in" and "out" relations ? "Le nord de Paris" is indeed ambiguous and can mean "a certain part located in Paris, in the north of it" as well as "a certain region out of Paris,

Figure 7: The south of a Bordeaux-Genève line

more in the north". Another problem concerns the socio-economic qualification of geographical areas. As already mentionned and shown in Table 1, such characterisations appear in a significant way in the corpus, and should be integrated in the matching process. Clearly again, simple word matching is not adequate and we must invoke some semantic knowledge, formalised as a semantic net or the like.

3.2 Computing a relevance degree

At this point, the system should be able to state if a given passage matches a query. However, a major task still remaining is to sort the passages from best to worst. This task is quite difficult since it involves heterogenous data.

3.2.1 Quantification

If we consider examples envolving quantitification, such as (4a)-(4d), we must admit that all entities which match the ZONE and TYPE need not to be relevant. For example in (4c) "some" means that only a part of the set of specified departments is concerned, on contrary to (4b) where "all" is the determiner. Which ones ? the linguistic expression does not say. Howeve we can compute a probability for any of these entities to be concerned so that we can propose the following ranked list: (4b), (4d), (4a), (4c) for a request concerning Calvados (department situated in the north of France):

(4a) The semantic of "the quarter" gives (no GIS needed) a wheight equals to 25%.

(4b) The semantic of "all" gives a wheight equal to 100%.

(4c) The semantic of "some" indicates that few entities are concerned. In this case, we stipulate a number of 5 entities (that's a heuristic). This leads to a wheight equals to $\frac{5}{n}$, n being the number of districts included in the zone. A request to the GIS gives $n = 52$. Hence, the weight is $\frac{5}{52} = 9.6\%$

(4d) In the same way, we obtain here $\frac{15}{52} = 29\%$.

Both linguistic knowledge (for semantic interpretation of the determiners) and geographical knowledge (for an evaluation of probability) will be needed in this process.

3.2.2 Granularity

We adopted a liberal strategy, which selects all passages *compatible* with the query. The counterpart is clearly the risk of noisy answers. The granularity criterion can provide a useful numeric evaluation of the relevance of an expression. For example, consider the query "find passages concening the city of Caen", with respect to passages mentionning "Caen" itself, "the Calvados" (department to which Caen belongs), "Basse-Normandie" (Caen's administrative Region), and finally "the northern half of France". For granularity reasons (left however for further consideration), it seems desirable to presents the four passages in this order, from the closest to the farthest level.

3.2.3 Negation

The paradox is as follows : if we say A is true for all X but Y, we positively say that not A is true for Y. Coming back to P2, the fact that Paris is explicitely excluded from the set of towns is a very strong information, almost as strong as a positive mention. We think this problem can be managed thanks to some *symetric*, i.e. negative relevance value. Hence P2 receives degree -1 with respect to query Q1. The user can choose to visualize or not passages with negative degrees among other answers.

4 Conclusion

The work presented in this paper concerns passage extraction from geographical documents. We focused on the most characteristic aspect of the project, namely the interpretation of the spatial component of texts and queries. We first described a linguistic analyser which performs a semantic analysis of expressions denoting geographical localisations. This analyser is operationnal and will in the future be developped and experimented, in order to cope with a greater variety of expression and cover large corpora. Then we addressed the question of the actual querying of text. We outlined a method for matching user requests with the computed representations of spatial expressions. We believe that some aspects of this process can be readily implemented. However it also raises some difficult questions, we plan to investigate in the next future.

Acknowledgements

Special thanks to Séverine Beaudet for her decisive contribution in the development of the spatial analyser.

References

Andrée Borillo, 1998, *L'espace et son expression en français*, Ophrys Press, Paris.

Frédérik Bilhaut, Thierry Charnois, Patrice Enjalbert, Yann Mathet, 2003, *Passag extraction in geographical documents*, Proc. Intelligent Information Systems 2003, New Trends in Intelligent Information Processing ans Web Mining, Zakopane, Poland (to appear).

James P. Callan, 1994, *Passage-Level Evidence in Document Retrieval*, Proc. 7th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland.

Michael A. Covington, 1994, *GULP 3.1: An Extention of Prolog for Unification-Based Grammar*, Research Report, AI.

Robert Hérin, Rémi Rouault, 1994, *Atlas de la France Scolaire de la Maternelle au Lycée*, RECLUS - La Documentation Française, Dynamiques du Territoire, 14.

Diego Molla, Rolf Schwitter, Michael Hess, Rachel Fournier, 2000, *Extrans, an Answer Extraction System*, Traitement Automatique des Langues, Hermes Science Publication, 41-2, 495-522.

Maria Teresa Pazienza (Ed.), 1997, *Information Extraction*, Springer Verlag.

Helmut Schmid, 1994, *Probabilistic Part-of-Speech Tagging Using Decision Trees*, Intl. Conference on New Methods in Language Processing. Manchester, UK.