

Natural Language Interfaces: What is the Problem? - A data-driven quantitative analysis

Philipp Cimiano¹ and Michael Minock²

¹WIS, TU Delft / ²University of Umea

Abstract. While qualitative analyses of the problems involved in building natural language interfaces (NLIs) have been available, a quantitative grounding in empirical data has been missing. We fill this gap by providing a quantitative analysis on the basis of the Geobase dataset. We hope that this analysis can guide further research in NLIs.

1 Introduction

So far, there has been an impressive amount of research on natural language interfaces (NLIs), i.e. on interfaces allowing users to interact with a certain information system in natural language. While NLIs are not inherently restricted only to the task of answering questions on the basis of a given database or knowledge base, most of the NLIs developed so far have been designed for this purpose. Along these lines, as in most other research on natural language interfaces, we limit ourselves to this restricted view of natural language interfaces essentially as systems providing answers to natural language questions in this paper. Research on NLIs dates back to the 70s and 80s (see [1], [6]) and has yielded increased attention in recent years with a plethora of systems emerging: PRECISE [13], STEP [11], ORAKEL [3], Aqualog [10], GINSENG [2], just to name a few of the very recent systems. What seems missing so far is a description of the problem, in particular a quantitative analysis of the problems inherent in the task of building natural language interfaces. While there have been qualitative analyses of the problems involved in constructing NLIs ([1], [6]), to our knowledge there has been no quantitative analysis grounding the qualitative characteristics of the problem in real data. This is crucial in our view as it can and should guide the development of NLIs in the future, focusing them on the challenging problems. It would also help system developers to focus on a specific phenomenon encountered in NLIs (e.g. resolution of ambiguities) and foster progress in the field by clearly designing and evaluating the solution to a specific phenomenon which would ideally not be specific to one particular approach but reusable across systems. In our view, no real progress can be expected in NLI research only from charts hiding the interesting details and solutions to characteristic problems involved in the task behind top performing precision and recall measures.

The structure of this paper is as follows: in the next Section 2 we describe the dataset we have used to provide a quantitative analysis and describe our methodology. Then, in Section 3 we describe our interesting findings and derive

conclusions in terms of requirements on NLI. In addition, we include some comments on portability (see Section 4) and on the issue whether deep syntactic and semantic processing is needed for an NLI (see Section 5). We conclude in Section 6 with a summary of our findings and implications for the development on NLIs.

2 Datasets and Methodology

To provide a quantitative analysis of the problem of constructing an NLI we proceeded as follows: we downloaded a dataset which has been frequently used for the evaluation of natural language interfaces, i.e. the Geobase dataset collected by Mooney and his students¹. The Geobase dataset describes states, cities, mountains, lakes, rivers and roads in the U.S., together with attributes such as area (state, lake), population (state, city), length (river), height (mountain, location) etc.

The datasets consists of a set of 880 test questions (actually 883 questions) and was collected through a web interface hosted at the University of Austin in Texas². We used the 883 test questions for our analysis. After downloading the dataset (in Prolog), we converted the whole dataset into the ontology languages F-Logic [9] and OWL³. The datasets are available from <http://www.cimiano.de> → Projects → Datasets and other Material → ORAKEL.

When converting the dataset into OWL and F-Logic, we used 7 concepts with a total of 17 different relations. We give below the concepts used together with their relations:

Concepts	Relations
state	name, abbreviation, capital, density, population, area, code hasCity, border, highest_point, lowest_point
city	name, area, inState
river	name, length, flowsThrough
mountain	name, inState, height
road	number, passesThrough
lake	name, area, inState
location	name, inState, height

The design above slightly deviates from the original schema in Mooney’s dataset, consisting of 8 relations (**state**, **city**, **river**, **border**, **highlow**, **mountain**, **road** and **lake**). We have essentially merged some of the information into one class (the class **state** thus containing the border as well as highest and lowest point information), removed some redundancies (e.g. the name of the state appearing in various relations) and added the **location** class which includes a **height** attribute for the location in question.

The original dataset of Mooney et al. consists of the following 7 relations:

¹ This dataset is available from: <http://www.cs.utexas.edu/users/ml/nldata.html>

² There is also a dataset consisting of 250 questions available from the University of Texas, but this is merely a subset of the larger dataset.

³ <http://www.w3.org/TR/owl-features/>

Relation	Attributes
state	name, abbreviation, capital, population, area, state_number, city1, city2, city3, city4
city	state, state_abbreviation, name, population
river	name, length, [states through which it flows]
border	state, state_abbreviation, [states that border it]
highlow	state, state_abbreviation, highest_point, highest_elevation, lowest_point, lowest_elevation
mountain	state, state_abbreviation, name, height
road	number, [states it passes through]
lake	name, area, [states it is in]

The main differences between our and the original dataset are the following:

- Some redundancies have been removed, keeping the name and abbreviation of a state only in the **state** class.
- The **border** and **highlow** relations have been removed and the information added to the **state** class.
- The height of the highest and lowest points has been modeled in the class **location**.
- The 4 cities in the **state** relation have not been modeled explicitly. We suppose these were the “major” cities in each state. In that case we can recover this information from the area/population information, thus defining “major” through one of these attributes rather than by its extension as in Mooney’s dataset.

For the 883 questions, we manually created F-Logic queries yielding the appropriate answers as result when evaluated with the OntoBroker [7] system, but also queries in generic logical form. All these settings do not restrict the general case in any way. Any other database or knowledge representation and query language could have been used in principle. The benefit of the F-Logic language as implemented by the OntoBroker system is that it provides built-in functionality for numerical comparisons as well as aggregation operators for calculating minima, maxima, sums etc., which, as we will see below, are crucial for the Geobase dataset.

We proceeded by manually annotating each of the questions together with characteristics that we regarded as relevant to our quantitative analysis. All annotations have been performed by one of the authors and suffer unavoidably from some subjective bias. The interesting findings obtained through these “annotations” are mentioned below. All cases are illustrated with example questions from Mooney’s dataset and the numbers directly refer to the questions as they are ordered in the dataset. The annotated dataset can be downloaded at <http://www.cimiano.de> → Projects → Datasets and other Material → ORAKEL.

3 Quantitative Analysis

3.1 Question Types

The questions were annotated with 4 question types:

- **wh-questions** (77.29%): the standard type of question starting with a *wh*-pronoun such as *What are major rivers in texas?* (201) or *What is the lowest point in texas?* (472)
- **how (adj/many) -questions** (17.38%), such as *How big is massachusetts?* (19) or *How many big cities are in pennsylvania?* (50).
- **requests** (3.59%): direct requests such as *Give me the cities in virginia.* (1) or *Can you tell me the capital of texas?* (5).
- **topicalized questions** (1.16%) where a certain entity is topicalized for the purpose of emphasis, as in *Iowa borders how many states?* (173) or a prepositional phrase is topicalized, as in *Of the states washed by the mississippi river which has the lowest point?* (186)
- **nominal** (0.58%): these are “questions” consisting of a single noun phrase, such as *people in boulder?* (187) or *rivers in new york* (189)

A system which therefore only supports standard *wh*-questions can only reach a recall of 77.56% on the Geobase dataset. This might partially explain the recall of significantly below 80% of the PRECISE system on this dataset (see [13]) as it requires that some ‘*wh-value*’ maps to a database element.

3.2 Language “light”

Our findings suggest that the language used in the questions is rather simple, containing a lot of ‘light syntactic constructions’ such as:

- copula verb ‘*be*’ (appearing in 59.43% of the sentences): e.g. *What are the high points of states surrounding mississippi?* (1), or *What is the major cities in montana?* [sic.] (482)
- light preposition ‘*of*’ (appearing in 21.52% of the sentences): *What are the highest points of all the states?* (210) or *What are the major cities of texas ?* (235)
- light verb ‘*have*’: (appearing in 12.46% of the sentences), e.g. *How many capitals does rhode island have ?* (51), or *How many rivers are in the state that has the most rivers?* (118)
- light preposition ‘*with*’ (appearing in 7.36% of the sentences): *How many people live in the state with the largest population density?* (104), or *What are the cities of the state with the highest point?* (209)

Thus, in many cases the relevant relations are not expressed directly in the text, but are hidden implicitly behind light constructions involving the verb “*to have*” as well as light prepositions such as ‘*with*’ and ‘*of*’. This is probably the reason why shallow approaches which ignore the linguistic details (for example ignoring non-content words in the input as in the PRECISE system), essentially relying on the structure of the knowledge base or data base to perform interpretation (such as PRECISE [13] or Aqualog [10]) are so successful on the type of questions the Geobase dataset consists of. Nevertheless, any NLI should implement techniques to deal with such a kind of lightweight (or semantically weak) constructions which require to infer the appropriate relation implicit in the surface realization.

3.3 Lexical ambiguities

While the problem of lexical ambiguities is mentioned in many overviews on natural language interfaces ([1], [6]), our findings suggest that classical lexical ambiguities are typically not the problem. For example, we did not find any single case of ‘classical’ noun homonymy or polysemy where the same word can convey completely different meanings in the dataset. Most of the ‘lexical’ ambiguities are actually introduced artificially by the knowledge base. For example, the adjectives ‘*big*’ and ‘*small*’ or the superlatives ‘*smallest*’ and ‘*largest*’ are ambiguous when modifying a state, as ‘size’ can be measured either in terms of area or inhabitants in the Geobase knowledge base. We found around 60 cases (6.80% of the questions) of such artificial ambiguities, e.g. *What is the largest state?* (422).

The fact that these cases are indeed ambiguous is corroborated by the fact that in some cases the users actually tried to disambiguate by specifying the property with respect to which size should be measured, as in (4 sentences, 0.45% of the questions): *What is the smallest state by area?* (579).

Further, there are further examples of reference ambiguities in the dataset e.g. *“How many people are there in new york?”* (77). In the latter case, it is not clear whether *New York* refers here to the city or to the state. There are many other and in some cases people do indeed try to disambiguate by adding the state for instance (in case of multiple cities with the same name in different states):

- What is the population of springfield missouri? (536)
- What is the population of springfield south dakota? (537)

In case no explicit disambiguation is provided, any NLI should request the user to disambiguate the input, in case of question 77 for example by asking: *“Do you mean New York city or the state of New York?”* or in cases like 422: *“Do you mean largest in terms of area or population?”*. The Aqualog system, for example, recognizes such reference ambiguities and asks a user for reference disambiguation.

3.4 Syntactic Ambiguities

Frequently, literature on NLI research also mentions syntactic ambiguities, especially attachment ambiguities, as problematic. To explore this phenomenon, we have annotated modifying prepositional phrases (PPs), relative clauses and modifying gerund constructions making explicit whether they i) attach to the only possible antecedent (noun phrase or verb phrase), ii) to the last one or iii) to a non-preceding constituent. In the case of prepositional phrases, we also distinguish the case of a PP providing essentially the predicate in a copula construct. We give examples for each of these cases below:

- PP attachment (only attachment point) (48.07%), i.e. *How many states in the us does the shortest river run through?* (166) or *Where is the highest point in hawaii?* (4)

- PP attachment (last attachment point) (40.62%), i.e. *How many people reside in utah?* (109) or *What is the capital of the state with the highest elevation?* (346)
- PP attachment (copula) (10.34%): *Tell me what cities are in texas?* (196), or *What mountains are in alaska?* (602)
- PP attachment (non-preceding attachment point) (0.48%): *What is the city in texas with the largest population?* (355) or *What is the state with the largest density in usa?* (588)
- relative clause (last attachment point) (72.92%): *What river traverses the state which borders the most states?* (612), or *What states border states that border colorado?*
- relative clause (only attachment point) (27.08%): *Give me the cities which are in texas?* (12) or *What is the smallest state that the mississippi river runs through?* (583)

So this means that by simply attaching PPs or relative clauses to the last constituent, we will take a correct decision in 99.27% of the cases for PP attachment (including last and only attachment point cases as well as the copula case where the PP functions as predicate) and in 100% of the cases for relative clauses (last and only attachment point). While we have not listed the gerund data explicitly, we get similar results, with last and only attachment points representing 100% of the cases. In other words, we found no gerund attaching to a non-preceding constituent.

As a consequence, a very simple baseline strategy which attaches every PP or relative clause to the last constituent will be difficult to beat. Thus, at least what the Geobase dataset is concerned, no substantial effort in syntactic disambiguation is needed. While this can not be claimed in the general case, we hypothesize that people try to produce less ambiguities when interacting with a natural language interface.

3.5 Scope ambiguities

Many natural language interfaces ignore scope ambiguities and even determiners (representing quantifiers) completely (e.g. PRECISE [13], Aqualog [10] etc.). In what follows we give examples for determiners/quantifiers appearing in the dataset:

- **question operator:** *What are major rivers in texas?* (201)
- **definites:** *What are the capitals of states that border missouri?* (250)
- **most:** *What is the capital of the state that borders the most states?* (344)
- **a:** *What is the largest city in a state that borders texas?* (407)
- **negation:** *How many rivers do not traverse the state with the capital albaney?* (124)
- **all:** *Show me all the major lakes in the us?* (193)
- **the least:** *What city has the least population?* (272)
- **at least:** *How many states border at least one other state?* (139)
- **fewest:** *Which rivers run through states with fewest cities?* (807)
- **each:** *What are the population densities of each us state?* (243)

Some statistics are given below:

Preposition	Occurrences (questions)	Rel. Percentage
in	280 (32.71%)	73.11%
through	100 (11.33%)	36.11%
next to	3 (0.34%)	0.78%
Total	383	100%

Operator	Occurrences (sent.)	Rel. Percentage
max	278 (31.48%)	55.71%
count	107 (12.12%)	21.44%
min	91 (10.31%)	18.24%
negation	12 (1.36%)	2.40%
sum	7 (0.79%)	1.40%
average	4 (0.45%)	0.89%
Total	499	100%

Table 1. Some statistics about spatial prepositions (left) and aggregation operators (right)

Scope taking elements	#Occurrences (sentences)	Rel. Percentage
question op. (all other than requests)	832 (94.22%)	51.81%
definites (incl. superlatives)	697 (78.94%)	43.40%
most	32 (3.62%)	1.99%
a	13 (1.47%)	0.81%
negation	12 (1.36%)	0.75%
all	11 (1.25%)	0.68%
least	7 (0.79%)	0.44%
fewest	1 (0.11%)	0.06%
each	1 (0.11%)	0.06%
Total	1606	100%

Clearly, there is a high occurrence of question operators, realized at the surface by *wh*-pronouns (not surprising), but there is also a very high number of definite noun phrases. However, it is important to mention that we have subsumed entity descriptions such as *‘the state of oregon’* under the category definites in the above table, but we have not taken them into account when calculating the number of scope-bearing elements in each question, which is almost 2 (Avg. 1.97, Std. Dev. 0.81), including plural NPs and excluding the uses of *‘the’* which are part of a named entity expression as mentioned above. The minimum number of scope bearing elements per sentence is 1 and the maximum 5. This shows that in principle there is one scoping decision to take per sentence. While in many cases it is true that this can be solved via some heuristics, i.e. *‘each’* outscopes every other quantifier in the question, definites are accommodated as high as possible, the question quantifier outscopes the rest of the quantifiers, etc. (compare the heuristics used in the TEAM system [8]), we are making a principled point here: strategies for disambiguation are needed, no matter how adhoc they are (as long as they are effective).

3.6 Spatial prepositions

While there is a significant number of light or vague prepositions in the dataset (see results in Section 3.2), spatial prepositions also tend to appear frequently, in particular the prepositions *in*, *next* and *through* (see Table 1 - left).

Examples are the following:

- **in:** *Name the rivers in arkansas.* (3), or *what is the biggest american city in a state with a river?* (301)
- **through:** *How many rivers run through texas?* (129), *what is the largest city in smallest state through which the mississippi runs?* (414)

- **next to:** *what states are next to arizona?* (694) *how many states are next to major rivers ?* (134)

The presence of spatial prepositions can be explained by the fact that the Geobase dataset is modeling locations as well as their spatial relationships. While some systems have shown that one can perform very well by essentially ignoring prepositions, we would like to make the point that the more principled solution would be to capture the domain-independent meaning of such spatial prepositions, allowing to reuse their meaning across domains. For example, ‘*in*’ has definitely a meaning in terms of spatial inclusion which is compatible with many domains if modeled appropriately, e.g. at the level of a foundational ontology as suggested in [4]. Taking into account the specific semantic contribution of spatial (and also temporal) prepositions gets important in many domains, especially in those including temporal knowledge (see Section 5).

3.7 Adjective Modifiers and Superlatives

There are at least 105 adjectives in the Geobase dataset (appearing in 11.89% of the questions) as well as 316 superlatives (appearing in 35.79% of the questions). The adjectives are distributed among the following cases: i) modifiers (6.9% of the questions) as in ‘*How many major cities are in arizona?*’ (67), ii) most/least+adj (0.91% of the sentences), as in ‘*What is the most populous city?*’ (487) and as iii) attribute selectors in how+adj questions (3.9% of the questions), as in *How big is alaska?* (18). Clearly, any NLI needs to handle adjective modification as well as superlatives. The challenge here is certainly that the interpretation of adjectives (and in consequence also of superlatives) is domain-specific and needs to be specified for each domain. For each adjective, an NLI needs information about the predicate in the knowledge base it represents (e.g. area/population in the case of small) as well as the polarity of the adjective, which is crucial to handle superlatives. It is also important to specify the conditions under which some object will fulfill the property represented by the adjective, e.g. specifying which is the minimum population for a city to be counted as ‘*big*’. An appropriate mechanism to provide such crucial and basic data about the meaning of adjectives is important to allow portability of NLIs. Many NLIs have ‘cheated’ in this respect, hardcoding the meaning of adjectives (e.g. ‘*major*’) and superlatives (e.g. ‘*shortest*’) in the backend system (see the evaluation engine in Prolog by Mooney where the appropriate ‘meaning’ has been hard-coded⁴). With respect to the PRECISE system, it is unclear at all how it can handle superlatives or modifying adjectives as they are clearly not semantically tractable, i.e. there is no column in the database corresponding to the superlative or to the adjective as modifier (requiring a specific value), so that additional mechanisms are needed beyond the algorithm described in [13] to handle adjectives and superlatives.

⁴ <ftp://ftp.cs.utexas.edu/pub/mooney/nl-ilp-data/geosystem/geoquery>

3.8 Aggregation, comparison and negation operators

We define aggregation operators as those calculating a minimum, maximum or a sum over a given set of values as well as those allowing us to count the number of individuals fulfilling a certain property. Comparison operators are those that allow to compare numbers w.r.t. to a given order (e.g. the one between integers).

Such aggregation operators are crucial to evaluate certain queries on the Geobase dataset. Before giving a few statistics, we will first give a few examples of questions requiring aggregation, comparison and negation operators. As we will see, such operators are “hidden” behind certain quantifiers and only appear in the translation into logical form:

- **counting**: e.g. *How many major cities are in arizona?* (67), requiring to count all those x which are major cities and are located in Arizona
- **maximum** (no counting involved): e.g. What cities in texas have the highest number of citizens? (269), returning those cities with a maximum number of inhabitants, where the number of inhabitants is given explicitly in the form of the population and does not need to be counted.
- **counting + maximum**: e.g. *What is the length of the river that runs through the most states?* (441), requiring to count, for each river x , the number of states that it flows through, taking the maximum over these and returning the length of the x flowing through the maximum number of states.
- **comparison**: e.g. *What states high point are higher than that of colorado?* (754), comparing the height of the high points of all states to the height of the highest point of colorado and returning those states with a higher point.
- **negation**: e.g. *Name the states which have no surrounding states?*
- **negation with counting**: *How many rivers do not traverse the state with the capital albany?* (124)
- **sum**: *What is the area of all the states combined?* (280)
- **average**: *Which state has the smallest average urban population?* (840)

It is clear that the occurrence of these ‘operators’ is highly correlated with the appearance of a certain quantifier in the surface form (see Section 3.5 and the corresponding analysis of quantifier frequencies). However, the challenge here is to predict how a certain quantifier will be realized as, i.e. as which set of logical operators. ‘How many’, for example can involve a counting operation or only a look-up, depending whether the information is modeled as a datatype property or an object property thus requiring to count objects standing in the relation in question:

- *How many capitals does rhode island have?* (51) (counting)
- *How many inhabitants does montgomery have?* (66) (no counting, only lookup)

The same holds for the quantifier most, also appearing in two different forms, one requiring a maximization only, the other one requiring a counting and a maximization operator:

- *What is the capital of the state with the most inhabitants?* (351) (max.)

- *What is the length of the river that runs through the most states?* (441)
(counting + max.)

This means that the correct interpretation of ‘*how many*’ and ‘*the most*’ depends on the way the relevant information has been modelled in the knowledge base (e.g. as datatype or object property). In other cases, the aggregator is in many cases only very loosely selected by the surface form of the question. In the sentence: ‘*What is the area of all the states combined?*’ (289), “all *x* combined” actually maps to a *sum* operator in the logical form but there is no reference to a sum in the surface form.). Some statistics about the occurrence of such “aggregators” are given in Table 1 (right).

3.9 Non-compositionality

In the Geobase dataset there is a high number of questions which with respect to their mapping into logical form are non-compositional, i.e. the logical form of the question is not exactly the composition of the meaning of the parts of the question. In particular, in many cases there are “parts” of the question which do not correspond to any element of the logical form. This is what we refer to as “non-compositionality”. According to our analysis, at least 12% (11.89%) of the questions in the Geobase dataset are non-compositional. Most of these questions include a reference to the USA which of course is not explicitly mentioned in the dataset as it only models information about the USA but never mentions this explicitly, as it is clear from the overall scope of the knowledge base.

A few examples of non-compositional elements in questions are given below:

- Give me the cities in virginia. (1)
- What are the major cities in the usa ? (232)
- What states in the united states have a city of springfield ? (755)
- What is the biggest american city in a state with a river? (301)

Of course, besides having elements of the query which do not appear in the logical form, we encounter also this situation the other way round, i.e. in many cases elements which appear in the logical form are not mentioned explicitly in the question. This is partly a byproduct of the light language used in the questions and essentially amounts to the cases we have discussed above, so that we do not analyze this any further here. So there is a need for NLI systems to ignore part of the input. This is for example accomplished by the ‘fudging’ operator in the STEP system [11].

3.10 Variability

One of the greatest challenges for any natural language interface is to handle the large variability in the way that a certain fact or relation can be expressed. It is certainly difficult to quantify variability and we will not even try to do so. The important observation is that the Geobase dataset is certainly no exception to the above. As an example, let us consider various forms in which one can ask for the population of a certain state:

- *How many inhabitants does montgomery have?* (66)
- *How big is the city of new york?* (23)
- *How many citizens in alabama?* (62)
- *How many residents live in texas?* (111)
- *How many citizens live in california?* (64)
- *How many people reside in utah?* (109)
- *What is the population of alaska?* (505)
- *population of boulder?* (188)

Any NLI should clearly somehow handle this variability, either by allowing people adapting the system to a new domain to encode it explicitly (as in ORAKEL [3] or STEP [11]) or implicitly in the way the question is mapped to a logical query by using the schema of the database or knowledge base.

3.11 Out of scope

In the Geobase dataset we find at least 17 (1.93%) questions which are definitely out of scope of the knowledge base. A few examples are:

- *How many states border the mississippi river?* (147): There is no information about which states border a river, but only about which rivers flow through a state.
- *What is the biggest state in continental us?* (317): There is no information about which states are on the continent and which not (e.g. Hawaii)
- *What is the length of the colorado river in texas?* (435): There is no information about the length of a river in a particular state (only the absolute length)
- *What is the maximum elevation of san francisco?*: No information about the highest points of cities is available (only for states).

A NLI should definitely inform the user in some way about the fact that the question is out of scope and not simply return no answer.

4 Portability: No Free Lunch!

While we have no data we could analyse with respect to portability issues, we think it is important to mention this issue as one of the most challenging problems in NLI development. Seldom have the efforts and resources needed to port a system been made explicit or compared accross systems (an exception being [12]). This would enhance our understanding of the problems and potential solutions to the portability problem. The most important issue here is to have some mechanism to specify how content words (verbs, nouns, adjectives etc.) map to predicates in the knowledge base. For sure, portability does not constitute a free lunch. ORAKEL and STEP rely on manual mappings created by a lexicon engineer on the basis of the data schema. Systems which seem to require no effort to be ported at first sight (e.g. [13]) require at least a lexicon which needs to be handcrafted or derived using general lexical resources such as WordNet in combination with the natural language labels available in the ontology or database schema. But WordNet clearly has limitations in scope, especially in

technical domains, so that porting to technical domains also comes at the cost of manually enhancing the lexica. Other systems which learn from training data require pairs of questions and queries in logical form [14], the provision of which constitutes a huge effort. The effort that is needed to customize an NLI to a new domain has been rarely quantified and compared across paradigms.

5 To deepen or not to deepen? That is the question.

As mentioned already, there have been many shallow approaches to NLIs with very impressive results on the Geobase dataset we have examined here (as well as other datasets collected by Mooney and colleagues). When looking at NLIs just from a narrow perspective, i.e. looking at their input (a question) and output (a formal query or the result of its evaluation), a shallow approach with almost 100% precision might look like the right solution. However, when expanding the capabilities of the NLI, we will encounter a number of obstacles when using a shallow system. We mention some of these capabilities below:

Paraphrase Generation: As argued recently by Minock [11] as well as others before [5], paraphrase generation is crucial to make sure that the system has captured the intended meaning of the user’s question. Paraphrase generation capabilities require some level of deeper representation (syntactic and semantic) to generate a non-ambiguous form of the question that a user can confirm or reject. Generating a non-ambiguous representation of the input sentence presupposes that the system is able to detect, represent, reason with and resolve ambiguities, which requires some level of deeper processing.

Discourse Processing: While most NLIs process each question in isolation (and they have been evaluated also in this mode), in real systems we expect that users will use pronouns to refer to previous entities and that they will provide fragmentary input (i.e. in the form of ellipsis). Any NLI should thus be designed in such a principled way that allows to extend it with some discourse processing capabilities. While the resolution of pronouns could be arguably carried out by pure statistical approaches neglecting discourse structure, the resolution of ellipsis would clearly benefit from some structural representation which allows to compute which gap the new content can fill. In any case, it seems that more sophisticated NLIs going beyond analyzing each sentence in isolation will require some deeper analysis and make explicit the discourse structure. This is also important when introducing new modalities such as gestures, prosody etc. While the integration of various modalities can also be approached in a “shallow” way (meaning statistical or adhoc), we claim that the integration could profit from an explicit and deep syntactic and semantic representation of the questions and their context.

Temporal Aspects: While many successful systems ignore non-content words such as determiners or prepositions, in some domains prepositions might indeed

matter. In the context of an ontology about american presidents, it makes a difference whether we ask for ‘*Who was president after WWII*’, vs. ‘*Who was president during WWII*’. In general, systems would thus profit from capturing the meaning of prepositions explicitly, possibly even assigning them a domain-independent meaning (as suggested in [4]).

Guidance: Guidance to a user in the task of formulating a query which is in line with the capabilities of the system is a very important feature of NLI. It requires, however, at least the existence of a grammar (in whatever form) which can be used to generate to suggest possible completions to the user. Ideally, the system would have one and only one grammar to be used in analysis and generation.

In summary, while it is tempting to explore shallow approaches which neglect deep linguistic and semantic structure, we think that this is the wrong way to go as it makes the extension of the system with respect to aspects mentioned above very hard, possibly requiring adhoc and principled extensions. For example, a system which ignores quantifiers in the semantic representation will have to integrate procedural attachments to simulate the ‘semantics’ of the quantifiers in an adhoc way. A system which ignores non-content words such as prepositions might have to incorporate prepositions in an adhoc manner when confronted with a domain in which they indeed matter (e.g. domains involving temporal aspects). Paraphrase generation and discourse processing are hard to achieve building on adhoc and shallow representations rather than on principled semantic representations. Building on sound theories of semantic representation and discourse would for example allow the import of all those insights on discourse processing from linguistics into our systems, while this “import” is not as straightforward when systems lack principled semantic representations. Overall, we think that building systems with deep syntactic and semantic processing capabilities will pay off in the long term, at least for the above mentioned reasons.

6 Conclusion and Future Work

We have started this research by wondering why there was no systematic quantitative analysis of the problem of providing NLI to databases or knowledge bases. While there have been many qualitative descriptions of the problems, a data-driven analysis has been missing so far. We have aimed at closing this gap. We can conclude that tackling the following issues is crucial for any natural language interface:

- handling other questions than wh-questions (otherwise we will miss about a quarter of the questions in the Geobaset dataset for instance)
- dealing with light syntactic/semantic constructions and vague prepositions (more than 60% of the questions contain such “light” constructions)
- including some approach for handling scope, in particular for representation and disambiguation (one disambiguation decision needed per question),

- generally, an adequate treatment of quantifiers, in particular those translated into aggregation, comparison and negation operators, which are very frequent in the dataset (over 40% of the questions involve some aggregation operator)
- accounting for the domain-specific meaning of adjectives (in various constructions) as well as superlatives
- treating prepositions (in particular spatial and temporal) in an appropriate way, ideally defining their domain-independent semantics
- handling non-compositional input
- handling and reacting appropriately to out-of-scope questions

Of course, to have a more substantial analysis, we would need to consider further datasets, e.g. Mooney’s restaurant and jobs datasets⁵. A first look at these datasets confirmed the prevalence of light language. Nevertheless, it also seems that these datasets do not involve as many aggregation operators, but contain mainly questions which can be answered directly from the database without any additional operations (by mere “look-up” so to speak). Nevertheless, considering more datasets could substantiate or not the conclusions derived from the analysis of the Geobase dataset. Nevertheless, we regard the Geobase as the most interesting one and it is also the one used and cited most frequently, so that we have started our analysis on the Geobase dataset. The fact that the jobs and restaurant domains are inherently “easier” is demonstrated by the higher performance in terms of recall on these datasets by PRECISE (see [13]).

As a next step, it would be interesting to consider in how far the different recent systems do indeed fulfill the above requirements. This would allow a meaningful qualitative comparison of the different systems and help to work out their strengths and weaknesses abstracting from the concrete results they yield on the dataset. Overall, we hope that with this analysis we will also be able to guide future research in NLI and help to focus on the challenging problems as well as to focus researchers on specific aspects of an NLI rather than trying to solve all problems with one single “black box” approach from which it is hard to tell how and why it solves the challenging issues mentioned in this paper and thus do not contribute much to our understanding of the task. With our analysis we also open the possibility for researchers to concentrate on a particular subset of the Geobase data to study a specific phenomenon (e.g. by selecting from our data the subset of questions with superlatives, the direct requests etc.) This allows exactly the type of research on a specific phenomenon that in our view is needed to push the field further. Past research has shown that we can get very good results with shallow approaches which neglect many of the difficulties we have mentioned here. However, we believe that such solutions will have i) problems to scale to more demanding domains (e.g. involving temporal knowledge) as well as ii) handle the small percentage of remaining questions which possibly require deeper processing (e.g. the PRECISE system, while being maximally precise, fails on more than 20% of the questions in the Geobase dataset).

On a more general note, while it is tempting to explore shallow approaches which neglect deep linguistic and semantic structure, we think that this is the

⁵ see <http://www.cs.utexas.edu/users/ml/nldata.html>

wrong way to go as it makes the extension of the system with “advanced capabilities” harder, possibly requiring adhoc and non-principled extensions. Overall, we think that building systems with deep syntactic and semantic processing capabilities will pay off in the long term.

Acknowledgements This research has been partly supported by the MULTIPLA project, funded by the Deutsche Forschungsgemeinschaft (DFG) under grant number 38457858.

References

1. I. Androutsopoulos, G.D. Ritchie, and P. Thanisch. Natural language interfaces to databases—an introduction. *Journal of Language Engineering*, 1(1):29–81, 1995.
2. A. Bernstein, E. Kaufmann, C. Kaiser, and C. Kiefer. Ginseng: A guided input natural language search engine for querying ontologies. In *Proceedings of the JENA User Conference*, 2006.
3. P. Cimiano, P. Haase, J. Heizmann, M. Mantel, and R. Studer. Towards portable natural language interfaces to knowledge bases: The case of the ORAKEL system. *Data and Knowledge Engineering (DKE)*, 62(2):325–354, 2007.
4. P. Cimiano and U. Reyle. Towards foundational semantics - ontological semantics revisited -. In *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS)*, volume 150, pages 51–62. IOS Press, 2006.
5. E.F. Codd. Seven steps to rendezvous with the casual user. In *Proceedings of the IFIP Working Conference on Data Base Management*, pages 179–200, 1974.
6. A. Copestake and K. Sparck Jones. Natural language interfaces to databases. *Knowledge Engineering Review*, 5(4):225–249, 1989. Special Issue on the Applications of Natural Language Processing Techniques.
7. S. Decker, M. Erdmann, D. Fensel, and R. Studer. Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. In *Database Semantics: Semantic Issues in Multimedia Systems*, pages 351–369. Kluwer, 1999.
8. B.J. Grosz, D.E. Appelt, P.A. Martin, and F.C.N. Pereira. Team: An experiment in the design of transportable natural language interfaces. *Artificial Intelligence*, 32:173–243, 1987.
9. M. Kifer, G. Lausen, and J. Wu. Logical foundations of object-oriented and frame-based languages. *Journal of the ACM*, 42:741–843, 1995.
10. V. Lopez and E. Motta. Ontology-driven question answering in aqualog. In *Proceedings of NLDB’04*, pages 89–102, 2004.
11. M. Minock. A phrasal approach to natural language interfaces over databases. In *Proceedings of NLDB’05*, pages 333–336, 2005.
12. M. Minock, P. Olofsson, and A. Naslund. Towards building robust natural language interfaces to databases. In *Proceedings of NLDB’08*, pages 187–198, 2008.
13. A. Popescu, O. Etzioni, and H. Kautz. Towards a theory of natural language interfaces to databases. In *Proceedings IUI’03*, pages 149–157, 2003.
14. C. Thompson, R. Mooney, and L. Tang. Learning to parse natural language database queries into logical form. In *Proceedings of the Workshop on Automata Induction, Grammatical Inference and Language Acquisition*, 1997.