[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland;
ed. Kirsten Falkedal (Geneva: ISSCO).]

Evaluation of Machine Translation Systems: A System Developer's Viewpoint

Jäppinen H., & Kulikov L.
SITRA KIELIKONE-project

1 Introduction

The evaluation of any artifact, system or event is – by definition – a subjective endeavor and tends to correlate with the value system of the evaluator. In an evaluation one attempts to make a decision about the significance or quality of the object under study, and that calls for a norm or a standard for comparison. There is great disagreement among people on these matters, and even for an individual, evaluation often depends on the intended use of the system or artifact.

Therefore, when one discusses the evaluation of Machine Translation systems, it should be clearly stated WHO evaluates WHAT and for WHAT PURPOSE. MT systems are commercial systems, and one can point out at least two clearly different value systems that have bearing on evaluation. Evaluation may have an impact on the sales of the system and it is natural for commercial companies to have evaluations published in the most positive light. On the other hand, in order to be able to estimate productivity changes caused by the system, a potential purchasing company or organization needs realistic or conservative evaluation estimates. Bearing this in mind, we can distinguish at least the following evaluation viewpoints:

- A potential user organization evaluates a system or systems vis-a-vis its/their effect on the organization (Albisser, 1991). This is the most comprehensive user viewpoint. It does not evaluate only productivity changes of individual translators but also the whole working environment. If translation is a vital function in an organization and if MT systems will be applied on a large scale, this viewpoint, which we may call ORGANIZATIONAL evaluation, is obviously highly recommendable. We have nothing further to say about this evaluation viewpoint.

- A potential user organization or a potential individual user evaluates a system based on its effects on the productivity of individual users. The evaluation attempts to estimate how translation throughput increases and how translation quality is affected if the workload is divided between a translator and an MT system. We call this viewpoint FUNCTIONAL evaluation.

- A potential user organization or a potential user or any other interested party evaluates the linguistic quality of an MT system. This viewpoint is the narrowest one: it is only interested in the quality of the translations the system is capable of producing automatically. We call this viewpoint LINGUISTIC evaluation. This perspective is of obvious theoretical interest, but it omits important pragmatic factors if not augmented with studies of the system in practical use.

- A commercial company evaluates a system in order to obtain a canned answer to potential customers when they ask: “Yes, but how good is it?” It is obvious that the publicly announced “facts” of the evaluations made by commercial companies are not reliable. A wish becomes too easily a “fact”. A discussion of this viewpoint would lead us into theory of persuasion and marketing.

- There is still one more viewpoint on evaluation, one that is of particular interest to us. The developer of an MT system evaluates the system in the course of its development in order to see what progress is being made and how the quality of the system improves correspondingly. The development of an MT system is a tough undertaking. There are many reasons why it takes years to build an MT system. One reason is that usually large lexicons have to be built from scratch. It is therefore important to obtain, as early as possible, hard quality estimates of the system in order to be able to make timely design choices.

We are in the final phases of finishing an MT system before delivering it to a customer for testing in actual use. We will discuss evaluation from the viewpoint mentioned last. Moreover, we will discuss the differences between functional and linguistic evaluation. But first we must briefly describe our system.

2 KIELIKONE MT Workstation concept

The SITRA Foundation in Finland is a public fund which allocates funds for projects of great national importance. In 1982 SITRA launched the KIELIKONE project for the purpose of designing computational models of the Finnish language. The short term goals were to obtain concrete language technology products; the simultaneous long term goal was to build an infrastructure for MT research. During its life-time so far the project has designed, implemented, and delivered to the market various software products for the Finnish language. A company has been established to market products and to carry on development work on them.

In 1986 a decision was reached that the project should now concentrate on full-scale MT research in cooperation with two major Finnish companies. One is the pilot customer for a Finnish-English system, which is now near completion. The other is the pilot customer for an English-Finnish system, a project in the somewhat remoter future.

The focus of our MT research has been the design of MT Workstations. By that term we mean personal computing systems which produce good quality raw translations and support post-editing with a user-friendly linguistic editor. There is no pre-editing phase. To promote wide applicability, the algorithmic part of the systems is language independent, and the parts holding language-dependent definitions are declarative.

These aims have been realized with the concept of an MT Machine, which holds the algorithmic part of any particular MT Workstation implementation. The MT Machine is a language-independent, general tree-manipulation system whose execution is controlled by a declarative rule base. The MT Machine itself is not confined to the use of any specific linguistic theory; our implementations commit us to the dependency theory as the model of deep sentence structure.

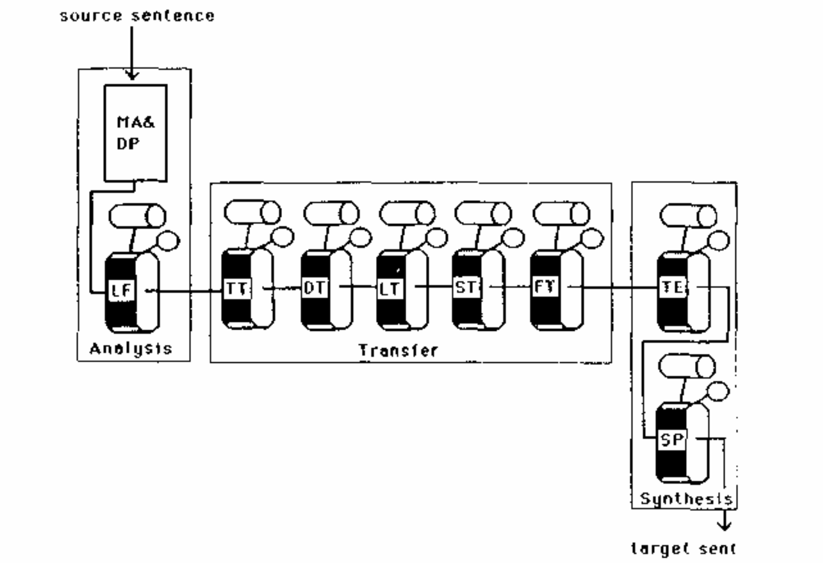

Figure 1. Linear architecture of the FE-implementation

The Finnish-English Workstation implementation has a simple linear architecture as shown in Figure 1.

A sentence is first subjected to morphological analysis and dependency parsing (MA & DP), and then follows a sequence of copies of the MT Machines, each copy having its own rule base. The analysis phase performs logical form reduction (LF), and then follow the five phases of the transfer part: term transfer (TT), domain specific lexical transfer (DT), general lexical transfer (LT), structural transfer (ST), and feature transfer (FT). The synthesis part expands the target logical form (TE) and produces the linear string of the target sentence (SP). Jäppinen et al. (1991) describes the system in more detail.

3 Interface

Interaction with the user takes place through a graphic interface. The screen is divided into input and output windows which display source language and target language sentences, respectively. The workstation concept takes post-editing seriously. One way of increasing translation quality in conjunction with positive user cooperation is to make editing and revising activities as convenient as possible. The user of the workstation can edit the text in the windows in different flexible ways. He/she can move text fragments around or delete or insert new words using services similar to those offered by modern text editors. If necessary, the user can also tag sentences for later scrutiny.

An important editing function is lexical replacement. It is a well known fact that one of the greatest problems in MT is the correct lexical choice. The rules of the MT Machine permit quite elaborate contextual checks in the lexical transfer phases. However, pragmatic knowledge, outside the text proper, often affects the translation of lexical items, and this knowledge is not within the reach of any finite rule system. The Finnish-English implementation features a dictionary of translation equivalents: Finnish words with sets of possible translation equivalents (in some contexts). If the user is not satisfied with a given lexical choice in the target text, he/she can point at the word, and a window with a list of alternative translations will appear on the screen. If an alternative is pointed at, it will automatically replace the wrong word in the text – even in the correctly inflected form.

Certain function keys are reserved for editing special problems arising from this particular language pair. For instance, Finnish does not have articles. Our workstation translates texts sentence by sentence. No attempt has been made to keep track of discourse referents. Consequently, no reliable method is available for deciding between definite or indefinite articles. We use heuristics (e.g. a subject normally has a definite article) and leave possible corrections for the user. When a specific function key is hit the article of the word under the cursor changes through the cycle of indefinite-definite-no article.

To summarize for the purpose of this discussion, the user has at his/her disposal function keys, the keyboard, and a mouse for editing. Using these facilities he/she can insert, delete, or replace individual words or sequences of words and reorder the word sequence of a given target sentence.

4 Testing and tuning

Anybody who tries to develop a comprehensive computational model of a natural language at the sentence level or higher faces the hard practical problem that all languages are vast and unbounded systems of communication. Probably there exists no published book which would depict, say, the syntactic structures of any natural language totally, without any residue. Yet, a computational system has to use a finite number of rules.

A belief in the generative nature of languages or at least their syntaxes offers a way out from this dilemma. A finite set of generative rules can model an infinite set of structures. An MT system has to utilize in the analysis phase a finite set of rules which reflect the generative nature of the source language syntax, Assuming that translation is a decomposable task, a finite set of translation rules will suffice for the translation part as well. A third vital component is a comprehensive bilingual lexicon. Due to the openness of languages it is clear that the rules and the lexicons are not immediately correct, but that they need elaborate testing and tuning against real data.

We have designed a systematic testing and tuning procedure for our Finnish-English implementation. We cannot go into details here. Basically, we pick, at random, pieces of text and translate them. Errors in different phases are analysed, recorded and corrected. Each of the phases has its own idiosyncratic error behavior. If we depict the sum of all errors we get what is represented in Figure 2.

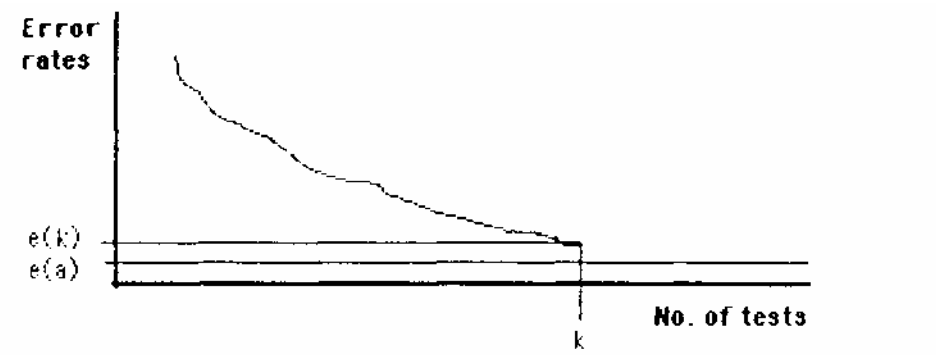

Figure 2. Error rates

Error rates (e.g. the number of errors per hundred sentences) should decrease steadily in the long run. So far, this has been our experience. However, there is no hope of reaching zero error rate within any finite time span. There will remain at least the lexical entries whose contextual checks do not take into account all possible occurrences of those words. Hence, there exists an asymptotic error rate (e(a)>0) which is the practical limit beyond which error rates decrease so slowly that further testing and tuning does not pay off. We have not yet reached the asymptotic error rate. (One should note that we mean linguistic errors in a fail-soft system. These errors are not fatal to the translation process; they only degrade the quality.)

Error rate analysis is an important tool from the viewpoint of software engineering. But it only tells us something about the internal state of the system. Undoubtedly, the most important external parameter of the system is translation quality. If we project the quality behavior of the system in the course of testing and tuning, we get Figure 3.

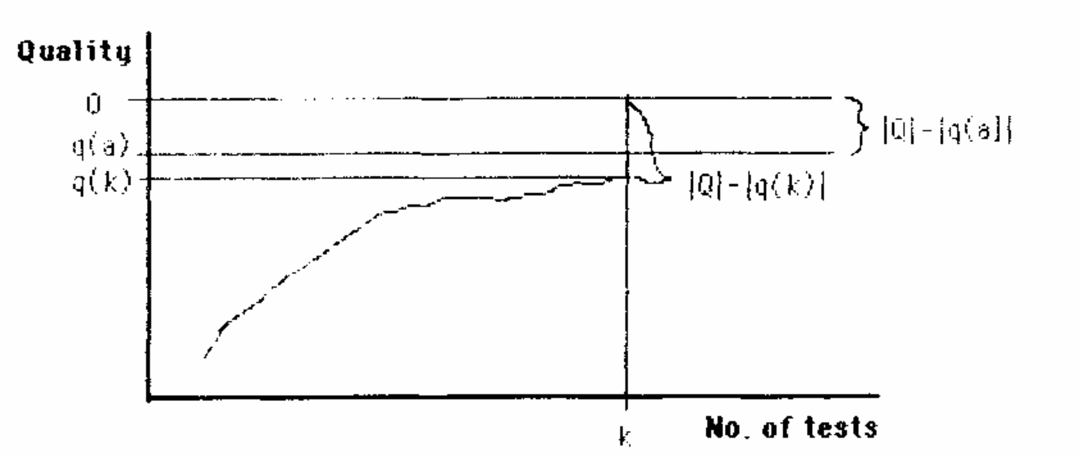

Figure 3. Quality improvement

For a given text corpus, there is a hypothetical upper limit quality parameter Q, which reflects the translation quality produced by a highly qualified human translator. Needless to say, we do not have any exact figure for Q. Another hypothetical upper limit q(a) (q(a)<Q) is produced by the system at its asymptotic error rate level. A third parameter, q(k) (q(k)<q(a)), tells us what the translation quality was in the test k.

We can observe the transitory state q(k) directly, but the stationary future state q(a) – which is more important – cannot be directly measured. However, an estimate q(a') of q(a) can be obtained by the following simple procedure. If we correct all general errors found in test k, and measure the quality after the corrections, it should represent a fairly good estimate of q(a).

For us, then, the quality of translation means precise figures for q(k) and q(a’) and a statistically reliable estimate of q(a) based on those figures. We cannot obtain absolute figures but we can get, as described below, relative figures Q-q(k), Q-q(a’) and Q-q(a) and let them stand for functional evaluation.

5 Functional vs. linguistic evaluation

So far in our work we have concentrated on the error rate analysis and have given only preliminary thought to the translation quality issue. First, there is the problem of a test domain. Should some specifically designed test sentences be used in an evaluation or should real data be used instead? The use of test suites has some clear advantages (King and Falkedal, 1991). They provide a good and objective view of translation quality when the evaluator is not aware of the internal structure of a system, or when several systems are being contrasted. But even then they should be supported by tests with real data. Since we are system developers, test suites are of no use to us. We would only get those answers we already knew. Only testing with real data, full of surprises, is a proper domain for us.

What, then, is a proper measure of quality for translation? It has been proposed that quality should be judged on the basis of the intelligibility of translations, on the one hand, and on their fidelity, on the other (Nagao, 1989). Such proposals emphasize linguistic evaluation over functional. The proposed parameters would indeed reflect the linguistic quality in the proper sense of the word. However, there are certain problems with these parameters. First, they exclude functional evaluation. Second, the measures rely on the inner workings of the human mind and are therefore subjective and difficult to measure in practice. Furthermore, such analysis, if carefully carried out, cannot be performed on large amounts of texts.

A more objective measure would refer to the external behavior of the test subjects. Instead of asking a competent translator questions about intelligibility and fidelity, which may be foreign to him/her, he/she is asked to do something natural. The most natural behavior vis-a-vis raw translations is correcting them. The quality measure of raw translations should therefore correlate with the editing operations performed. The fingers of the translator do the talking, so to speak. In terms of Figure 3, editing operations “reduce” the distance between the raw translation and an acceptable one, and a measure of functional quality can be expressed as the “distance” between the two translations (similarly for q(k)):

(1) $Q\text{-}q(a') = C_1 \text{ x Sum-of}(o_l) + C_2 \text{ x Sum-of}(o_2) + \ldots C_3 \text{ x Sum-of}(o_n)$

where o_i is an application of the editing operation of type $_i$
C_i is the weight assigned to the operation of type $_i$

(1) represents a view of functional evaluation, and it measures linguistic quality only indirectly. Besides being objective, such behavioral criteria have additional benefits. It is often the case that a less than perfect translation suffices. The behavioral criteria are easily adjusted to relative quality requirements. If lower quality suffices, the translator performs fewer editing operations, and the system gets a higher rating. Linguistic evaluation measures, such as intelligibility and fidelity, do not render relative figures so easily.

We have not yet fully developed the idea of behavioral criteria. Some open questions remain, such as should we observe “higher” level behavior by counting different high level editing operations, possibly weighting them as proposed in (1), or should we rely on more elementary operations and weight them equally? Presently we are experimenting with the simplest operations -- keyboard strokes and mouse clicks – and a proposed estimated evaluation “distance” for the system at its asymptotic error level is the one shown in (2).

(2) $$Q\text{-}q(a') = \text{Sum-of}(o_1) + \text{Sum-of}(o_2)$$

where o_1 is an application of a keyboard stroke
o_2 is an application of a mouse click.

If (2) is divided by the number of characters in the text, we get a normalized figure (3), which we may call editing intensity (EI), as the relation of the editing work to a manual translation. A similar suggestion appears in Brown et al. (1990).

(3) $$EI = Q\text{-}q(a') / NC$$

where NC is the number of characters in the text.

Editing time rate (ETR) is another normalized functional quality figure. By ETR we mean the time spent on editing in relation to the manual translation of similar text (4). A standard for Finnish-English translations used by professional translators is 1560 characters/hour.

(4) $$ETR = Te \times 1560 / NC \times 3600$$

where Te is the editing time in seconds.

An obvious disadvantage of behavioral criteria is their weak correlation with linguistic quality. For a given pair of sentences the evaluation measures (2) and (3) only occasionally reflect the linguistic quality of the sentences. It is easy to produce two translations of one sentence which differ so that their translation quality in the linguistic sense is quite different (say, the one is ungrammatical and the other requires just stylistic changes) and yet their values (3) are identical. The behavioral criteria cannot be applied to individual sentences or to a small set of sentences. They are only applicable to statistically significant amounts of texts.

6 Closing remarks

We want to close this discussion with an example. The appendix includes two extracts of original Finnish news items as they appeared in a major Finnish newspaper and their raw translations produced by the current version of our Finnish-English system after those errors had been omitted which could be corrected in a general fashion. No ad hoc corrections have been made. Hence, the raw translations estimate the system's capabilities at the asymptotic level (a'). Also shown are the post-edited translations. Notice that the system capitalizes unknown words and uses proper name as the default syntactic category. For these texts we get the following estimates for the editing intensity and the editing time rate.

$$\begin{aligned} \mathrm{EI1} &= (173 + 84) / 1017 &&= 0.25 \\ \mathrm{ETR1} &= 721 \times 1560 / 1017 \times 3600 &&= 0.31 \end{aligned}$$

$$\begin{aligned} \mathrm{EI2} &= (293 + 139) / 1648 &&= 0.26 \\ \mathrm{ETR2} &= 956 \times 1560 / 1648 \times 3600 &&= 0.25 \end{aligned}$$

The significance of these evaluations is debatable. The system has not yet reached the asymptotic level, and the two small pieces do not represent a statistically significant amount of text. On the other hand, there was not sufficient training in the use of this editor prior to running these tests and the editing times are probably higher than they would be in real use. One should also note that the persons post-editing the texts were not professional translators and no precise quality requirements were stated for them.

Bearing these reservations in mind, the figures quite unanimously suggest a productivity increase by a factor of about four. One might argue that the post-edited translations do not represent high enough quality and prove wanting. On the other hand, we are not performing fine-tuning yet and therefore the estimated asymptotic level is probably lower than the level will be after sufficient amount of testing and tuning ($q(a')<q(a)$). It seems fairly safe to assume that the system would increase the productivity of this text type at least by a factor of two to three at the asymptotic error level.

References

Albisser, D., Evaluation of MT Systems at Union Bank of Switzerland, this volume.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P., *A statistical approach to machine translation.* Computational Linguistics, Vol. 16, No. 2, June 1990.

Jäppinen, H., Hartonen, K., Kulikov, L., Nykänen, A., and Ylä-Rotiala, A., *KIELIKONE MT Workstation.* Proceedings of MT Summit III, Washington D.C., July, 1991.

King, M. and Falkedal, K., *Using test suites in evaluation of machine translation systems.* Proceedings COLING-90, 1990, 211-216.

Nagao, M., *Machine Translation: How Far Can It Go?,* Oxford University Press, 1989.

Appendix

Example 1

- uk3_51.src

Saarinen neuvoo verolippuaan yhä odottelevia ihmisiä ottamaan yhteyttä verovirastoon. Jos kadonneita verolippuja ei löydetä, puuttuvat verotiedot aiotaan joko tulostaa uudelleen tai pahimmassa tapauksessa laskea käsin. Joka tapauksessa tietonsa voi käydä tarkistamassa veronäyttelyssä. Kaikkien ilmoitusten välissä on laskelma, josta voi katsoa miten verotuksessa kvi. Saarinen korostaa, että verotus on kaikkien osalta valmis ja palautuksetkin tulevat ajallaan, vaikka dokumentti siitä puuttuukin. Saarisen käsityksen mukaan kenenkään ei myskään tarvitse olla huolissaan mahdollisista maksamatta jäävistä lisäveroista. Kaikki verolipuitta jääneet ihmiset saavat ilmeisesti palautusta. Uudet lomakkeet postitettiin verovelvollisille torstaina pikapostissa. Uskomme, että uudet lomakkeet tulevat perille maanantaina, kertoo toimistopäällikkö Martti Eveli. Evelin arvion mukaan myöhässä lomakkeensa saavia ihmisiä on noin tuhat.

- uk3_51.raw

Saarinen advices people waiting for their demand note still to contact Inland Revenue. If the missing demand notes are not found, will either to output the missing tax information again or to calculate by hand in the worst case be intended. In any case they can go to the tax exhibition to check its information. Between all announcements, there is a calculation from which one can look how in taxation it went. Saarinen emphasizes that taxation for everybody is ready, and the returns will come in due course even though the document about it is missing, too. In the opinion of Saarinen, nobody also needs to be worried about the possible surtaxes that will not be payed either. All people that missed the demand notes obviously get a return. The new forms were mailed on Thursday to the taxpayers by express mail. We believe that the new forms will arrive on Monday, the office manager EVELI says. According to the estimate of EVELI, there are about a thousand people getting their form late.

- uk3_51.edited

Saarinen advices people who are still waiting for their demand note to contact Inland Revenue. If the missing demand notes are not found either the missing tax information will be printed again or in the worst case calculated by hand. In any case people can go to the tax exhibition to check their information. Between all announcements there is a calculation from which one can see how taxation went. Saarinen emphasizes that taxation for everybody is ready and the returns will come in due course even if the documents about them are missing. In the opinion of Saarinen nobody needs to be worried about the possible surtaxes that remain unpaid. All people who missed the demand notes obviously get a return. The new forms were mailed on Thursday to the taxpayers by express mail. We believe that the new forms will arrive on Monday the office manager EVELI says. According to the estimate of EVELI there are about thousand people getting their form late.

- uk3_51.statistics

```
Thu Apr 30 13:19:10 GMT-2:00 1992
uk3_51
Characters, at beginning: 1038 at end: 1017
Words: 167 Sentences: 10
Key clicks: 173 Button clicks: 84 Sum: 257
Time elapsed: 721 seconds
```

Example 2

- ut3_61.src

Yhdysvaltain telakkateollisuusyhdistys vaatii kongressilta lainsäädäntöä, jonka tarkoituksena on estää valtion tukiaisten turvin rakennettujen laivojen käynnit amerikkalaisissa satamissa. Lainsäädäntöhanke yritettiin saada liikkeelle edustajainhuoneessa aiemmin tässä kuussa. Hankkeen takana on laivanrakentajien neuvosto, joka haluaa esityksellään ennen muuta vauhdittaa paikalleen juuttuneita neuvotteluja telakkatukiaisista teollisuusmaiden järjestössä. Järjestön piti jo viime vuoden maaliskuun loppuun mennessä päästä sopimukseen telakkatukiaisten poistamisesta. Neuvottelut ovat kuitenkin jumiutuneet ja subventioista on hyvää vauhtia tulossa uusi kauppapoliittinen riitakapula toisaalta Yhdysvaltain ja toisaalta Euroopan yhteisön, Japanin ja Etelä-Korean välille. Tukiaisriidan kiristyminen johtuu myös kansainvälisen kauppajärjestön neuvottelujen vaikeuksista. Ne ovat tulehduttaneet kauppapoliittista ilmapiiriä ja nostattaneet protektionistisia tunnelmia nimenomaan Yhdysvaltain kongressissa. Tällaisessa ilmapiirissä on vahvasti uudelleen pintaan noussut myös vanha kiista lentokoneteollisuuden tukiaisista. Neuvosto on esittänyt edustajainhuoneelle lakia, joka määräisi erityisen sakkomaksun Yhdysvaltojen satamassa käyvälle alukselle, mikäli sen rakentamisessa on käytetty valtion tukiaisia. Maksu olisi niin suuri, että se tyrehdyttäisi tällaisten alusten liikenteen Yhdysvaltoihin.

- ut3_61.raw

The **TELAKKATEOLLISUUSYHDISTYS** of the United States requires from the congress a legislation whose purpose is to prevent in the American harbours the visits of the ships built with the help of the state subsidies. An attempt was made to get a bill moving earlier this month in the House of Representatives. Supporting the plan, there is the shipbuilders' council which above all wants to speed up in the organization of the industrial countries negotiations on the shipbuilding subsidies that got stuck with its presentation.

The organization had to agree by the end of March last year on the removal of the shipbuilding subsidies. The negotiations, however, have come to a deadlock and the subventions are becoming a new politico-commercial bone of contention between on the other hand the United States and on the other hand the European Community the Japan and the South Korea quickly. The tightening of the quarrel of subsidies is caused also by the difficulties in the negotiations of the international trade organization. They have worsened a politico-commercial atmosphere and have raised protectionist atmospheres particularly in the congress of the United States. In such an atmosphere, also the old dispute over the subsidies of the aeroplane industry has risen strongly again to the surface. The council has presented to the House of Representatives a bill which would determine the special fine to the vessel visiting the harbour of the United States if state subsidies have been used in its building. The payment would be so big that it would stop the traffic of such vessels to the United States.

- ut3_61.edited

The Association of shipbuilding industry in the United States requires from the congress a legislation whose purpose is to prevent ships built with the help of the state subsidies from visiting the American harbours. Earlier this month an attempt was made to get a bill moving in the House of Representatives. Supporting the plan there is the shipbuilders' council which above all wants to speed up negotiations on the shipbuilding subsidies that got stuck with their presentation in the organization of the industrial countries. By the end of March last year the organization had to agree on the removal of the shipbuilding subsidies. The negotiations have however come to a deadlock and the subventions are quickly becoming a new politico-commercial bone of contention between on the other hand the United States and on the other hand the European Community Japan and South Korea. The tightening of the quarrel of subsidies is also caused by the difficulties in the negotiations of the international trade organization. They have worsened the politico-commercial atmosphere and raised a protectionist atmosphere particularly in the congress of the United States. In such an atmosphere the old dispute over the subsidies of aeroplane industry has also risen strongly to the surface. The council has presented the House of Representatives a bill which would determine a special fine to a vessel visiting the harbour of the United States if state subsidies have been used in building the vessel. The payment would be so big that it would stop the traffic of such vessels to the United States.

- ut3_61.statistics

```
Thu Apr 30 13:50:52 GMT-2:00 1992
ut3_61
Characters, at beginning: 1670 at end: 1648
Words: 257 Sentences: 10
Key clicks: 293 Button clicks: 139 Sum: 432
Time elapsed: 956 seconds
```